VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

DIPLOMOVÁ PRÁCE

Brno, 2020 Bc. Marek Mikulec

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV TELEKOMUNIKACÍ
DEPARTMENT OF TELECOMMUNICATIONS

SYSTÉM ZABEZPEČENÉHO PŘENOSU A ZPRACOVÁNÍ DAT Z AKTIGRAFU
SYSTEM OF SECURED ACTIGRAPH DATA TRANSFER AND PROCESSING

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. Marek Mikulec**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Jiří Mekyska, Ph.D.**
SUPERVISOR

BRNO 2020

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Diplomová práce

magisterský navazující studijní obor **Informační bezpečnost**

Ústav telekomunikací

Student: Bc. Marek Mikulec ***ID:*** 185936
Ročník: 2 ***Akademický rok:*** 2019/20

NÁZEV TÉMATU:

Systém zabezpečeného přenosu a zpracování dat z aktigrafu

POKYNY PRO VYPRACOVÁNÍ:

Cílem práce je návrh a implementace zabezpečeného systému, který umožní vyčítat data z aktigrafu GeneActiv a odesílat na server, kde bude možné k datům přistupovat a zpracovávat je metodami strojového učení. Bude vytvořeno webové rozhraní, které bude uživatel používat k monitorování dat v čase a k analýze časových řad odpovídajících fázi spánku (např. identifikace časových oken, ve kterých uživatel spal). Jelikož bude systém přenášet, uchovávat a zpracovávat citlivá data pacientů, bude kladen velký důraz na zabezpečení a příslušnou legislativní úpravu. V případě použití data setu třetí strany bude analyzována legislativní úprava regulující využití dat pro medicínské účely.

DOPORUČENÁ LITERATURA:

[1] VAN HEES, Vincent T., Séverine SABIA, Kirstie N. ANDERSON, et al. A Novel, Open Access Method to Assess Sleep Duration Using a Wrist-Worn Accelerometer. PLOS ONE. 2015, 10(11). DOI: 10.1371/journal.pone.0142533. ISSN 1932-6203.

[2] GÉRON, Aurélien. Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. Tokyo: O'Reilly, [2017]. ISBN 978-1-4919-6229-9.

Termín zadání: 3.2.2020 ***Termín odevzdání:*** 1.6.2020

Vedoucí práce: Ing. Jiří Mekyska, Ph.D.

prof. Ing. Jiří Mišurec, CSc.
předseda oborové rady

ABSTRAKT

Nový koncept Health 4.0 přináší myšlenku spojení moderních technologií z oblasti vědy a techniky s výzkumem ve zdravotnictví. Tato práce realizuje v duchu konceptu Health 4.0 systém zabezpečeného přenosu a zpracování dat z aktigrafu GENEActiv. Systém je úspěšně navržen, implementován, otestován a zabezpečen. S pomocí neinvazivní metody monitorování pohybu a teploty subjektu pomocí aktigrafu GENEActiv umožňuje systém bezpečným způsobem přenést, zpracovat a vyhodnotit data o spánkovém okně subjektu pomocí algoritmu strojového učení XGBoost. Navržený systém je v souladu s platným právem České republiky a splňuje zákonné požadavky.

KLÍČOVÁ SLOVA

Aktigrafie, zabezpečený přenos dat, strojové učení, XGBoost, monitorování spánku, Health 4.0

ABSTRACT

The new Health 4.0 concept brings the idea of combining modern technologies from field of science and technology with research in healthcare and medicine. This work realizes a system of secured actigraph data transfer and preprocessing based on the concept of Health 4.0. The system is successfully designed, implemented, tested and secured. With the help of a non-invasive method of monitoring the movement and temperature of the subject using the GENEActiv actigraph allows the system to securely transfer, process and evaluate the subject's sleep data using the machine learning algorithm XGBoost. The proposed system is in accordance with the valid law of the Czech Republic and meets legal requirements.

KEYWORDS

Actigraphy, secured actigraph data transfer, machine learning, XGBoost, sleep monitoring, Health 4.0

MIKULEC, Marek. *Systém zabezpečeného přenosu a zpracování dat z aktigrafu*. Brno, 2020, 95 s. Diplomová práce. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací. Vedoucí práce: Ing. Jiří Mekyska, Ph.D.

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma „Systém zabezpečeného přenosu a zpracování dat z aktigrafu“ jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

podpis autora

PODĚKOVÁNÍ

Rád bych poděkoval vedoucímu diplomové práce panu Ing. Jiřímu Mekyskovi, Ph.D. za odborné vedení a milou a přívětivou spolupráci. Chtěl bych poděkovat své přítelkyni, rodině a kamarádům za pomoc a podporu při tvorbě této diplomové práce a při studiu obecně.

Brno ....

podpis autora

Obsah

Seznam obrázků

Seznam tabulek

Seznam výpisů

Úvod

Jedním z fenoménů dnešní doby je tzv. Health 4.0. Jedná se o aplikaci moderních technologií z oblasti vědy a techniky do odvětví zdravotnictví a medicínského výzkumu. V duchu myšlenky Health 4.0 přináší práce možnost spojit aktigrafii s metodami strojového učení.

Aktigrafie je neinvazivní metoda sledování subjektů, která umožňuje dlouhodobě zaznamenávat pohyby. Často se využívá k vyšetřování poruch spánku, neboť na základě získaných dat lze mimo jiné odhalit časové okno spánku, případně může být možné určit spánkové fáze a nezvyklé patologické jevy [1].

Strojové učení představuje označení pro soubor metod, které na základě poskytnutých dat přináší nové poznatky s využitím výpočetní techniky. Aniž by byl počítač k dané činnosti explicitně naprogramován, dokáže díky poskytnutým datům například rozpoznat souvislosti, nalézt anomálie a podobně [2].

Spojení těchto dvou metod přináší nové možnosti uplatnění v rámci monitorování spánku. Aby však vše mohlo fungovat, musí být navržen a implementován systém, který umožní získat data pomocí aktigrafu a shromáždít a předzpracovat je k následné aplikaci metod strojového učení. Metody strojového učení poté musí být vhodným způsobem aplikovány, aby na základě dat z aktigrafu dokázaly vyhodnotit výsledky jako například spánkové okno.

Cílem této diplomové práce je navrhnout a implementovat systém, který umožní extrakci dat z aktigrafu a jejich následný bezpečný přenos na centralizované webové úložiště na kterém budou data zpracována metodami strojového učení. Práce si rovněž klade za cíl prozkoumat stav techniky a prověřit příslušnou legislativní úpravu, aby výsledný systém byl v souladu s platným zákonem České republiky.

Diplomová práce je členěna do sedmi kapitol. První kapitola představuje stav techniky a zhodnocení existujících řešení pro možnosti aktigrafie. Výstupem je volba aktigrafu, který bude v rámci systému použit. Druhá kapitola prezentuje teoretický návrh systému a rozdělení procesních rolí. Třetí kapitola se zabývá otázkou zpracování získaných dat, představuje měřené veličiny a teoreticky představuje metody strojového učení. Ve čtvrté kapitole je podrobně popsáno navržené a realizované řešení systému. Pátá kapitola detailně popisuje implementaci zpracování dat pomocí metod strojového učení. Šestá kapitola představuje proces a výsledky testování realizovaného řešení a ověřuje funkčnost systému. V rámci sedmé kapitoly se práce zaměří na legislativní úpravu související s realizovaným systémem a prověří splnění zákonem stanovených podmínek pro praktické použití systému.

1 Stav techniky

V této kapitole se práce věnuje průzkumu stavu techniky. Výstupem kapitoly je výběr vhodného již existujícího hardwarového a softwarového řešení umožňujícího pokročilé monitorování životních funkcí subjektu.

1.1 Specializovaná řešení – aktigrafy

Aktigrafie je metoda, která využívá specializovaná přenosná zařízení zaznamenávající v delším časovém úseku pohyb subjektu. Aktigrafie je pro výzkum spánku využívána již delší dobu, existují dokonce i standardizované metodiky pro její správnou aplikaci [1, s. 519–529].[1]

1.1.1 mbientlab

Jednou ze společností, která se zabývá výrobou aktigrafů je společnost mbientlab. Senzory 9-axis IMU umožňují sběr informací o pohybu, okolním osvětlení, teplotě a podobně. Tyto informace se přenáší na zařízení se systémy Android, iOS, Linux či Windows. K datům se dá přistoupit přes API, které podporuje programovací jazyky `Swift`, `JavaScript`, `C#` a další. Zařízení má výdrž 2 dny až 2 týdny a paměť 8 Mb. Cena zařízení je přibližně 100 $ ≐ 2220 Kč.[2]

1.1.2 ActiGraph

Aktuálním modelem společnosti nese označení wGT3X-BT. Toto zařízení je kompatibilní se systémy Android a iOS. Umožňuje monitorovat pohyb, tep, okolní osvětlení a podobně. Zařízení wGT3X-BT má výdrž 25 dní, paměť 4 GB a voděodolnost do jednoho metru. K datům se dá přistupovat pomocí systému společnosti ActiGraph. Cena není explicitně stanovena, odvíjí se od domluvy s výrobcem a velikosti objednávky, standadně se pohybuje kolem 180 $ ≐ 4000 Kč za kus.[3] Zařízení bylo použito v řadě odborných studií [3, s. 13–20].

[1] Aktigrafy se často využívají pro studie spánku, jedná se o diskutované téma, viz `https://www.researchgate.net/post/What_are_your_preferred_actigraphy_devices`.

[2] Informace jsou aktuální k 21. 9. 2018 a pocházejí ze stránek výrobce `https://mbientlab.com/research-development-kits`.

[3] Informace jsou aktuální k 21. 9. 2018 a pocházejí ze stránek výrobce `https://www.actigraphcorp.com/activity-monitor-comparison`.

1.1.3 Activinsights

Společnost vyrábí například zařízení GENEActiv. Tento aktigraf má výdrž baterie závislou od frekvence záznamů dat, která je nastavitelná. Například při frekvenci 100 Hz je výdrž baterie 7 dní, při frekvenci 10 Hz až 45 dní. Zařízení je voděodolné, cena zařízení se pohybuje kolem 200 $ ≐ 4400 Kč. Data lze ze zařízení jednoduše extrahovat pomocí aplikace GENEActivPcSoftware dostupné pro OS Windows. Dále mohou být data zpracována libovolným způsobem.[4]

1.1.4 Philips Respironics

Jedná se o dceřinou společnost korporátu Philips zabývající se vývojem bio-senzorů. Aktuálním modelem jsou hodinky Actiwatch Spectrum PRO. Ty umožňují nahrávat data o spánku, pohybu, okolním osvětlení, čí získávat data přímo z manuálních záznamů subjektů. Hodinky umožňují záznam na baterii 60 dnů, a záznam dat bez synchronizace po dobu 22 dnů. Odolnost náramku proti vodě je IP52. Data jsou zpřístupněna skrze software výrobce.[5] Cena se pohybuje kolem 2000 $ ≐ 44000 Kč. Zařízení bylo využito v rámci jiných studií [3, s. 13–20].

1.2 Nespecializované řešení – wearables

Wearables je souhrné označení pro spotřební elektroniku, která je určená k monitorování sportovních výkonů a životních funkcí především neprofesionálních sportovců. Zařízení se specializují na prezentování naměřených dat pomocí srozumitelných aplikací, data jsou ve velké míře filtrována a agregována již na daném hardware.

1.2.1 Fitbit

Společnost Fitbit je renomovaným výrobcem wearables. Škála výrobků je rozdělena na zařízení, která jsou vybavena Fitbit OS, a ta která jej neobsahují.

Posledním zařízením bez Fitbit OS je náramek Charge 3. Ten vydrží na baterii 7 dní, velikost paměti ani frekvence synchronizování není uvedena. Náramek je voděodolný.[6] Lze přistupovat pouze k datům upraveným na samotném zařízení, což

[4] Informace jsou aktuální k 23. 9. 2018 a pocházejí ze stránek výrobce `https://www.activinsights.com/`.

[5] Informace jsou aktuální k 23. 9. 2018 a pocházejí ze stránek výrobce `http://www.actigraphy.com/solutions/actiwatch/actiwatch-pro-specifications.html`.

[6] Informace jsou aktuální k 23. 9. 2018 a pocházejí ze stránek výrobce `https://www.fitbit.com/au/charge3`.

jsou podrobná data o ušlé vzdálenosti, tepové frekvenci a spánkových fázích. K čistým datům z akcelerometru se přistupovat nedá, z toho důvodu nelze využít toto zařízení jako aktigraf.[7] Cena náramku je kolem 4000 Kč.

Zařízení s Fitbit OS je například model Versa. Ten má výdrž baterie 4 dny a není voděodolný. Kromě dat přístupných i z Charge 3 umožňuje také omezený přístup k datům z akcelerometru, s tím že limit paměti pro aplikaci je 15 Mb.[8] Zařízení lze tedy s omezeními využít jako aktigraf. Cena zařízení je přibližně 300 $ ≐ 6600 Kč.

1.2.2 Mobvoi

Ticwatch S a E od společnosti Mobvoi jsou hodinky s operačním systémem Android Wear OS. Hodinky vydrží na baterii 2 dny, velikost paměti je 4 GB, hodinky jsou voděodolné IP67. Díky systému Wear OS je možné napsat aplikaci pro získávání dat z akcelerometru, a tak lze zařízení využít jako aktigraf. Hodinky samy zaznamenávají informace o srdečním tepu a krocích. Cena je přibližně 130 $ ≐ 3000 Kč.[9]

1.2.3 Misfit

Misfit Vapor Smartwatch jsou vybaveny operačním systémem Android Wear OS. Výdrž baterie je 2 dny, velikost paměti 4 GB, voděodolnost do 50 metrů. Poskytují stejné možnosti jako Ticwatch, cena je 200 $ ≐ 4400 Kč.[10]

1.2.4 Garmin

Hodinky Garmin Vivoactive3 umožňují zaznamenávat informace o spánku, ušlé vzdálenosti a řadu dalších funkcí, které mohou být zakoupeny volitelně. Výdrž baterie je až 7 dnů. K datům z akcelerometru se dá přistoupit skrze platformu connect-iq, zařízení lze tedy použít i jako aktigraf. Hodinky jsou voděodolné, velikost paměti není uvedena. Cena hodinek je 7500 Kč.[11]

[7] Informace pochází z diskuze na fóru Fitbit `https://community.fitbit.com/t5/Web-API-Development/Get-the-raw-accelerometer-data/m-p/388498`.

[8] Informace pochází z diskuzí na fóru Fitbit `https://community.fitbit.com/t5/Web-API-Development/Get-the-raw-accelerometer-data/m-p/388498` `https://community.fitbit.com/t5/SDK-Development/Memory-capacity-and-usage/td-p/2248186` `https://dev.fitbit.com/build/reference/device-api/accelerometer/`.

[9] Informace jsou aktuální k 23. 9. 2018 a pocházejí ze stránek výrobce `https://www.mobvoi.com/eu/pages/ticwatchse`.

[10] Informace jsou aktuální k 23. 9. 2018 a pocházejí ze stránek výrobce `https://misfit.com/misfit-vapor`.

[11] Informace jsou aktuální k 23. 9. 2018 a pocházejí ze stránek výrobce `https://www.garmin.cz/garmin-vivoactive3-optic-silver-white-band/78947` a ze stránek pro vývojáře `https://developer.garmin.com/connect-iq/compatible-devices/`.

1.2.5 Samsung

Zařízení Samsung Fit2 Pro umožňuje zaznamenávat informace o ušlé vzdálenosti, má výdrž baterie asi 3 dny, velikost paměti 4 GB a je voděodolné. Nativně neposkytuje informace o spánku a nemá senzor měření tepové frekvence. Zařízení obsahuje operační systém Tizen, který umožňuje přístup k datům akcelerometru s pomocí vlastní aplikace. Cena zařízení je 5000 Kč.[12]

1.3 Shrnutí

Byl proveden průzkum stavu techniky. Nabízí se více alternativních řešení, která lze k implementaci využít. Řešení byla shrnuta v rámci tabulky 1.1.

Tab. 1.1: Přehled zařízení pro záznam spánku

Firma	**Model**	**Výdrž baterie**	**Paměť**	**Vodě -odolnost**	**Data z akcelero -metru**	**Senzor tepu**	**Cena**
mbientlab	9-axis IMU	7 dnů	8 Mb	IP40	Ano	Ne	2 200 Kč
ActiGraph	wGT3X-BT	25 dnů	4 Gb	1 m	Ano	Ano	4 000 Kč
Activinsights	GENEActiv	45 dnů	4 Gb	1 m	Ano	Ne	4 400 Kč
Philips Respironic	Actiwatch Spectrum Pro	60 dnů	32 Mb	IP57	Ano	Ne	44 000 Kč
Fitbit	Charge 3	7 dnů	Neznámá	50 m	Ne	Ano	4 000 Kč
Fitbit	Versa	4 dny	15 Mb	Ne	Ano	Ano	6 600 Kč
Mobvoi	Ticwatch E	2 dny	4 Gb	IP67	Ano	Ano	3 000 Kč
Misfit	Vapor	2 dny	4 Gb	50 m	Ano	Ano	4 400 Kč
Garmin	Vivoactive3	7 dnů	Neznámá	5 ATM	Ano	Ano	7 500 Kč
Samsung	Fit2 Pro	3 dny	4 Gb	5 ATM	Ano	Ano	5 000 Kč

1.4 Zvolené řešení

K řešení diplomové práce byl zvolen aktigraf od společnosti Activinsights, náramek GENEActiv. Důvodem byla pořizovací cena, univerzální využití, neboť zařízení disponuje výstupem v podobě binárních nekomprimovaných dat, dostupná dokumentace, množství odborných prací využívajících toto zařízení [4] a dobré výsledky v porovnání s plně profesionální technikou vyšší cenové kategorie [5, s. 13–15].

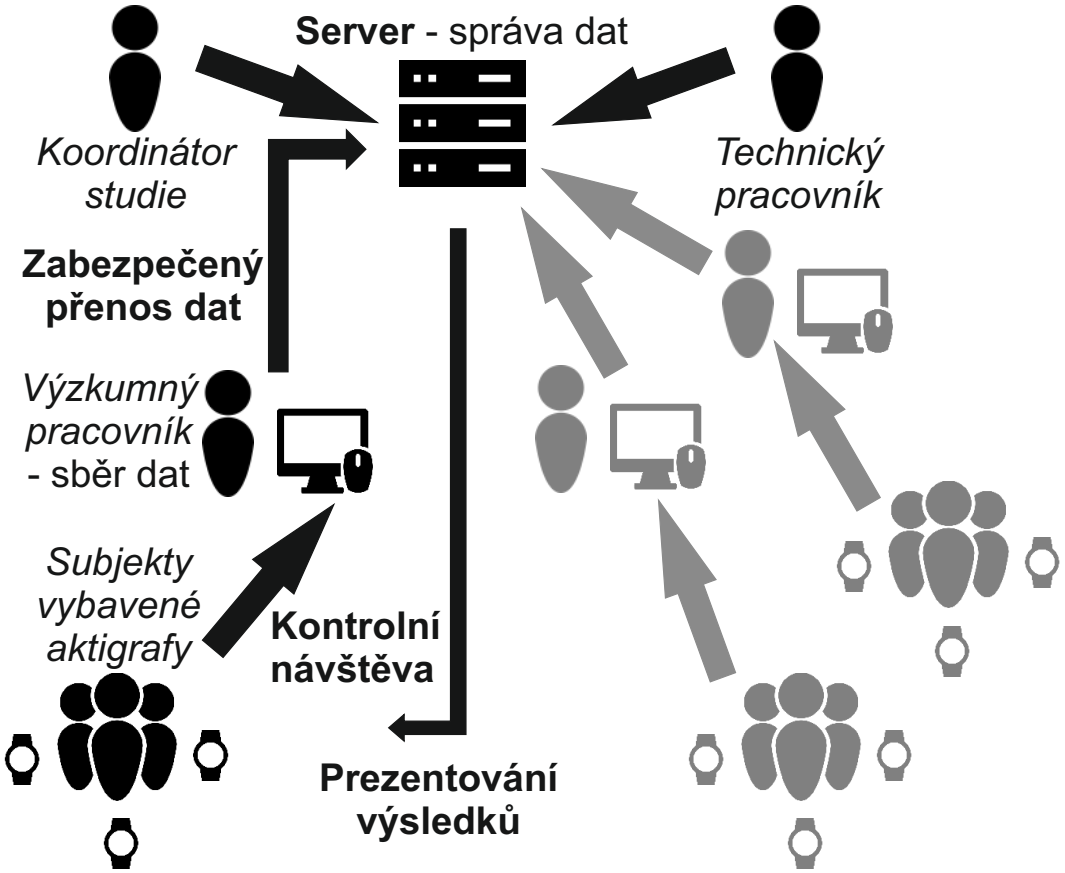
[12] Informace jsou aktuální k 23. 9. 2018 a pocházejí ze stránek výrobce `https://www.samsung.com/cz/wearables/gear-fit2-pro/SM-R365NZKAXEZ/` a ze stránek pro vývojáře `https://developer.tizen.org/community/tip-tech/obtaining-device-orientation-and-acceleration-using-built-sensors`.

2 Návrh systému

Data nasbíraná náramkem budou v rámci diplomové práce dále zpracována. K tomuto účelu byl navržen systém, který umožňuje agregovat a zpracovat data a také prezentovat výsledky. Bude představen návrh systému zpracování dat, definovány procesy v rámci systému a určeny zodpovědnosti jednotlivých subjektů.

2.1 Systém zpracování dat

Systém lze nejlépe prezentovat následujícím schématem 2.1, které rovněž popisuje hlavní způsob použití systému a transakce dat skrze něj.



Obr. 2.1: Schéma systému[1]

Jak je ze schématu patrné, data budou shromažďována od subjektů, které budou aktigraf používat po předem stanovenou dobu. Poté proběhne kontrolní návštěva, při které výzkumný pracovník stáhne data z náramku a převede data z binární

[1]Ikony byly vytvořeny autory Freepik a Google a staženy ze stránky `www.flaticon.com`.

podoby do formátu CSV pomocí aplikace GENEActivPCSoftware. Následně data nahraje na centrální server zabezpečeným kanálem. Na serveru proběhne zpracování dat metodami strojového učení. Poté bude možné na webovém rozhraní serveru prozkoumat výsledky jednotlivých měření. V celém systému bude zajištěna vysoká míra bezpečnosti dat.

2.2 Rozdělení zodpovědnosti subjektů

Aby byl systém plně funkční, je třeba stanovit zodpovědnost v rámci procesu sběru a analýzy dat. K tomuto účelu byla sestavena takzvaná RACI matice zodpovědnosti za jednotlivé aktivity. V této matici jsou jednotlivým procesním rolím přiděleny následující míry zodpovědnosti:

- **R = responsible** – odpovědnost za vykonání dané aktivity,
- **A = accontable** – odpovědnost za vykonání aktivity tak, jak je předdefinována, většinou přidělena vedoucí osobě,
- **C = consulted** – podílí se na vykonání aktivity, ale nepřebírá za ní zodpovědnost,
- **I = informed** – jde o roli, která musí být o výsledku či výstupu aktivity informována [9].

Tab. 2.1: RACI matice systému

	Vyšetřované subjekty	Výzkumný pracovník	Technický pracovník	Koordinátor studie
Konfigurace aktigrafu	I	R		A
Sběr dat	R	A		
Kontrolní návštěva	R	A		I
Stažení dat z aktigrafu	I	R		
Předzpracování dat		R	A	A
Přenos dat na server		R	A	
Konfigurace serveru		I	R	A
Tvorba uživatelských účtů		I	R	A
Předzpracování a tagging dat			I	R
Návrh modelu strojového učení			I	R
Implementace a tréning modelu			R	A
Aplikace naučeného modelu			R	I
Interpretace výsledků		I	R	R
Publikování výsledků	I	I	R	A

3 Zpracování dat

Je důležité představit měřené veličiny a formát dat. Poté budou prezentovány a popsány metody pro další zpracování dat.

3.1 Měřené veličiny

Data poskytovaná náramkem jsou vyjmenována v tab. 3.1.

Tab. 3.1: Data poskytovaná náramkem GENEActiv

Veličina	**Jednotka**	**Rozsah dat**	**Rozlišení**
Zrychlení v ose X	g	$(-8, 8)$	$0,0039\,\mathrm{g}$
Zrychlení v ose Y	g	$(-8, 8)$	$0,0039\,\mathrm{g}$
Zrychlení v ose Z	g	$(-8, 8)$	$0,0039\,\mathrm{g}$
Úroveň osvětlení	lux	$(0, 5000)$	$5\,\mathrm{lux}$
Stav tlačítka	on/off	$\{0, 1\}$	
Teplota	°C	$(0, 70)$	$0,1\,^{\circ}\mathrm{C}$

Za povšimnutí stojí jednotka g používaná pro měření akcelerace, tzv. „g-force“. Nejedná se o základní jednotku zrychlení dle SI, tou je $\mathrm{m} \cdot \mathrm{s}^{-2}$. Převod je následující: $1\,\mathrm{g} = 9,80665\,\mathrm{m} \cdot \mathrm{s}^{-2}$ [7, s. 26].

3.2 Formát dat

Náramek GENEActiv poskytuje data ve formě časových řad. Tato data jsou z náramku extrahována jako binární `.bin` soubory pomocí nativní aplikace GENEActivPC-Software, dodávané společně s aktigrafem [7, s. 19]. Binární soubory mohou být zpracovány přímo cílovou aplikací, například pomocí open source knihovny `Pampro` v jazyce `Python`.[1] Při pokusné implementaci se v knihovně vyskytovaly problémy s verzí jazyka `Python`, a tak bylo potřeba některé metody upravit přímo v kódu knihovny `Pampro`. Extrakce dat byla i přesto velmi pomalá a nestabilní.[2]

Proto bylo následně využito funkce aplikace GENEActivPCSoftware, která umožňuje převést data z binárního souboru do souboru `.csv` – Comma-separated

[1] Knihovna je dostupná na odkaze `https://github.com/Thomite/pampro`. Knihovna umožňuje kromě extrakce dat z binárních souborů společnosti Activeinsight také extrakci dat z aktigrafů jiných výrobců [8].

[2] Pokusná implementace je dostupná na odkaze `https://gitlab.com/MarekMikulec/geneactiv-processing-data/tree/bit-read-save`. Informace je aktuální ke dni 3. 3. 2019.

values [7, s. 20]. Tento soubor pak lze načíst buď jako dataset aplikace Excel, případně lze načíst pomocí vhodného skriptu do paměti cílové aplikace provádějící zpracování dat. [3]

3.3 Strojové učení

Strojové učení je vědecká metoda, která umožňuje počítači získat znalosti na základě dat bez explicitního naprogramování. Datová sada, která se k danému účelu využívá, se obvykle označuje termínem *dataset*. Strojové učení použijeme obvykle v případě, kdy bychom potřebovali k řešení úkolu definovat velké množství různých podmínek a pravidel, které nemusí být vůbec zřejmé. S další dávkou dat se pravidla navíc mohou zcela změnit. Strojové učení se snaží ve zpracovaných datech objevit právě taková pravidla a souvislosti, která jsou rozumnou aproximací řešení [2, s. 20–24].

3.3.1 Rozdělení metod strojového učení

Strojové učení lze rozdělit dle řady různých parametrů. Následuje výčet základních metod dělení.

Rozdělení dle lidského vstupu

Na základě míry lidského vstupu do procesu učení rozdělujeme strojové učení na učení s učitelem, bez učitele apod.

Strojové učení s učitelem bude prezentováno na následujícím příkladu: učitel označí data výsledkem, který od algoritmu očekává. Následně poskytne algoritmu velké množství takto označených dat a očekává, že se naučí data rozdělovat dle poskytnutého značení [2, s. 26–27].

Strojové učení bez učitele oproti tomu nepracuje s předem označenými daty. Systém musí vyvodit závěry bez vstupu učitele. Výsledkem algoritmu může být tedy například rozdělení dat do logických skupin (*cluster*), nalezení souvislostí mezi skupinami a tak dále [2, s. 27–30].

Kombinací strojového učení s učitelem a bez učitele, tzv. *semisupervised learning*, můžeme chápat jako přístup, kdy část dat označí učitel a zároveň část dat zůstává bez značení. Některé informace o datech poskytne učitel, další si vyvodí systém sám. Jedná se o hojně využívanou metodu [2, s. 31].

[3] Skript pro načtení dat z CSV souboru do datových struktur jazyka Python je dostupný na odkaze `https://gitlab.com/MarekMikulec/geneactiv-processing-data/tree/csv-to-python-save`. Později byl tento skript pozměněn a použit v rámci práce pro vykreslení grafu.

Dalším přístupem je tzv. *reinforcement learning*. Počítač má určitou sadu úkonů, které muže odsimulovat, a na základě situace která poté nastane, dostává zpětné hodnocení. Tento přístup ovšem není pro práci relevantní [2, s. 31]

Rozdělení dle okamžiku učení modelu

Algoritmy strojového učení dále klasifikujeme na základě míry, s jakou se přizpůsobují nově příchozím datům.

Strojové učení po dávkách, *batch learning*, využívá celý dataset rozdělelný na jednotlivé dávky k natrénování modelu před jeho nasazením do produkčního prostředí. Nově příchozí data jsou klasifikována natrénovaným modelem, ovšem model se z těchto dat neučí. Inkrementální učení probíhá na principu doplnění původního datasetu o nová data a opětovném natrénování modelu pomocí všech dostupných dat. Aktualizovaný model je následně vyměněn za model původní. Jelikož tréning datového modelu na celém dostupném datasetu je časově i výpočetně náročný, neprovádí se aktualizace v reálném čase [2, s. 33].

Stojové učení v reálném čase, *online learning*, využívá příchozích dat k přeučení existujícího modelu v reálném čase. Data jsou rozdělena po miniaturních dávkách, učení na těchto malých datasetech probíhá rychle a může být prováděno v reálném čase. Tento přístup je vhodný pro souvislý tok dat, na která je třeba reagovat co nejrychleji. Metoda může rovněž pomoci s ušetřením úložného prostoru, neboť již zpracovaná data, která byla jednak klasifikována a jednak použita jako vstup pro učení modelu, nemusí být dále uložena. Velkou výzvou je správné nastavení míry učení, *learning rate*. Pokud nastavíme příliš nízkou míru učení, bude model reagovat pomalu na probíhající změny, ovšem bude více imunní vůči šumu v datech. Vysoká míra učení naopak přinese rychlé reakce, ale je zde vyšší citlivost na případný šum [2, s. 33–34].

Rozdělení dle generalizace

Dalším způsobem, jak rozdělit metody strojového učení, je dle míry generalizace.

Učení založené na instanci, *instance-based learning*, využívá známé případy. U nových případů měří míru podobnosti s již známými případy a na základě této míry generalizuje.

Učení založené na modelu, *model-based learning*, vytváří ze známých příkladů model. Na základě modelu se snaží systém provést předpovědi pro nová data. Cílem je namodelovat predikční metodu, která bude dosahovat korektních výsledků pro nová data. Správnost je vyjádřena fitness funkcí nebo cenou. V případě fitness funkce se snažíme dosáhnout co nejvyšších hodnot, v případě ceny co nejnižších [2, s. 36–39].

3.3.2 Proces dolování znalostí

Proces strojového učení se snaží z poskytnutých dat dolovat znalosti. Dle zdroje [6] proces sestává z následujících sedmi kroků.

1. Čištění dat.
    - Odstranění nekvalitních dat a nekonzistentních dat.
2. Integrace dat.
    - Kombinace několika datových vstupů.
    - Je třeba sjednotit formát dat, např. mohou být použity odlišné jednotky.
3. Výběr dat.
    - Vybíráme data z databáze, které chceme analyzovat.
4. Transformace dat – *preprocessing*.
    - Konsolidace či transformace dat do podoby vhodné pro dolování znalostí.
5. Dolování znalostí.
    - Aplikace učících se algoritmů.
6. Ohodnocení naučeného modelu.
    - Posouzení kvality modelu na základě objektivních kritérií.
7. Reprezentace znalosti.
    - Vizualizace naučené znalosti uživateli, či jiný výstup [6, s. 167].

4 Implementace systému

Tato kapitola se zabývá praktickou realizací systému. Zodpovězeny budou otázky „Jak a s pomocí čeho byl systém vytvořen?“ a budou uvedeny problémy, které bylo třeba vyřešit při implementaci.

4.1 Hardwarové vybavení

Ke sběru dat byl využit výše avizovaný aktigraf GENEActiv od společnosti Activinsights. Systém je připraven specificky pro zpracování dat ve formátu poskytovaném tímto náramkem. Parametry náramku byly specifikovány v kapitole 1.3.

Pro serverovou část byl použit herní notebook Acer Aspire V15 Nitro Black Edition kvůli absenci specializovaného hardware. V reálné implementaci systému by bylo nutné použít dostatečně výkonný server s adekvátním množstvím paměti, aby byla zajištěna bezproblémová obsluha klientských požadavků. Parametry notebooku jsou uvedeny v příloze A.

Pro stažení dat z náramku lze využít jakýkoliv počítač s operačním systémem Windows, který splňuje požadavky aplikace GENEActivPCSoftware, viz manuál výrobce [7]. Nahrát a následně analyzovat data lze z jakéhokoliv zařízení pomocí webového prohlížeče, který přistupuje k webovému rozhraní serveru.

V rámci celého systému tak vyjma aktigrafu nejsou požadavky na specializované hardwarové vybavení a lze jej provozovat s poměrně malými náklady na pořízení i provoz zařízení.

4.2 Implementace serverové části

Serverová část je klíčovým bodem celého systému. Stará se o shromáždění a analýzu dat, celkovou správu studie, prezentování výsledků a podobně. Jedná se také o kritický bod bezpečnosti celého systému.

4.2.1 Operační systém

Serverová implementace byla realizována ve virtualizovaném prostředí operačního systému Fedora. K virtualizaci byl použit nejprve program VMware Workstation 15, později pak kvůli problémům se zobrazením UI byl virtuální počítač migrován do programu Oracle VM Virtualbox. Vše bylo spuštěno pod operačním systémem Windows 10.

Virtualizace operačního systému byla zvolena z důvodu bezpečnosti, potenciální útočník by neměl mít možnost poškodit hlavní operační systém Windows ani

v případě kompromitace virtualizováného OS. Taktéž je díky virtualizaci možné jednoduše zálohovat celý systém. Usnadněna je rovněž potencionální migrace systému na jiný server, kdy lze přenést celý obraz operačního systému a spustit jej na jiném stroji nezávisle na odlišnosti hardware. Navíc lze přidělovat hardwarové prostředky dle potřeby. Nevýhodou je pak vyšší náročnost, kdy virtualizovaný operační systém nemůže plně využít všechny dostupné prostředky systému. Ty je třeba rozdělit mezi hostitelský a virtualizovaný operační systém.

Operační systém Fedora byl zvolen s ohledem na vysokou míru bezpečnosti a důslednou péči o aktuálnost systému ze strany společnosti RedHat i ze strany komunity uživatelů.[1] Systém podporuje množství kvalitních vývojářských nástrojů. S přihlédnutím k použitému programovacímu jazyku `Python` a frameworku `django` je rovněž vývoj na Linuxových distribucích považován za komfortnější v porovnáním s operačním systémem Windows.

4.2.2 Programovací jazyk

Hlavní programovací jazyk projektu je jazyk `Python`. Ten byl zvolen na základě velkého množství kvalitních knihoven především pro oblast strojového učení.

Pro vyšší bezpečnost řešení nebyl v rámci systému využit vestavěný systémový interpret jazyka `Python`, který řídí důležité funkce operačního systému Fedora, ale bylo využito virtuálního prostředí poskytovaného nástrojem `Conda`. Tento nástroj taktéž umožňuje přívětivější správu balíčků jazyka `Python`.[2]

Byl použit programovací jazyk `Python` ve verzi 3.7. Ten je v současnosti považován za standard pro vývoj aplikací s využitím strojového učení. Vhodný je taktéž pro serverovou část, neboť podporuje balíčky specializovaných frameworků pro vývoj webových aplikací.

4.2.3 Webové frameworky

Pro usnadnění vývoje webového rozhraní serveru je vhodné využít existující framework. Byly prozkoumány různé webové zdroje za účelem volby správného frameworku.[3] Volba probíhala následně především mezi frameworky `flask` a `django`.

[1] Bližší informace o operačním systému Fedora založeném na Linuxovém jádře lze získat na adrese `https://getfedora.org`. Verze 31 byla ke dni 2. 11. 2019 nejaktuálnější verzí, systém byl po celou dobu studie pravidelně aktualizován pro zajištění bezpečnosti a stability.

[2] V rámci práce byla použita nejaktuálnější verze nástroje `Conda`, například k 3. 11. 2019 verze 4.7.12. Více informací o nástroji `Conda` lze získat na domovské stránce `https://docs.conda.io/en/latest/`

[3] Informace brané v potaz pro volbu frameworku vycházely z následujících webových stránek: `https://hackr.io/blog/python-frameworks`, `https://medium.com/fintechexplained/flask-host-your-python-machine-learning-model-on-web-b598151886d` – podrobný návod

Framework `flask` je mikro-framework, který poskytuje základní nástroje pro tvorbu webových aplikací. Je snadno rozšiřitelný, lze do něj přidat všechny potřebné komponenty a zároveň nepoužít komponenty redundantní.

Framework `django` poskytuje připravené komplexní řešení, obsahuje nástroje pro správu všech vrstev webové aplikace od databáze a správy uživatelských účtů až po frontendové nástroje. Jedná se také o masivně používaný framework s kvalitní dokumentací a množstvím tutoriálů. Velký důraz je rovněž kladen na bezpečnost. Dle *Open Web Application Security Project – nadace zaměřená na bezpečnost webových aplikací* (OWASP) pochází většina bezpečnostních chyb pouze ze špatné konfigurace uživatele [10]. Z výše uvedených důvodů byl pro implementaci zvolen framework `django`.

4.2.4 Webový server

Webový server byl spuštěn jako služba pod přihlášením uživatele bez administrátorských práv, aby byla zajištěna vyšší bezpečnost. Využita byla implementace webového serveru integrovaná v rámci frameworku `django`. Při nasazení na reálný webový server [11] lze aplikaci integrovat s komerčně využívanými serverovými implementacemi, například se server `Apache` či `NGINX`.

Server byl zpřístupněn pouze v rámci privátní sítě a nebyl připojen do internetu, jelikož se z důvodu plynoucích z bezpečnostních opatření nařízených během epidemie Covid19 nemohla uskutečnit pilotní studie.

Server byl zabezpečen technologií `https`, celá komunikace probíhá po ustanovení bezpečné relace šifrovaně. K zajištění funkcionality protokolu `https` je potřeba disponovat platným certifikátem podepsaným důvěryhodnou autoritou. V rámci diplomové práce tak byly vytvořeny dva certifikáty pomocí knihovny `OpenSSL`. První certifikát simuloval certifikační autoritu a umožňoval vystavovat další certifikáty. Certifikát byl importován do úložiště důvěryhodných certifikačních autorit testovacího systému. Druhý certifikát byl poté vystaven pro samotnou webovou stránku a podepsán pomocí certifikátu certifikační autority. Následně již byla veškerá komunikace zajištěna pomocí protokolu `https`.

V praktickém nasazení v rámci sítě internet by pro webovou stránku bylo zakoupeno doménové jméno a vystaven certifikát, který by byl podepsán reálnou certifikační autoritou. Všechny tyto služby jsou ovšem zpoplatněny a v rámci diplomové práce nebylo třeba je provádět. Byla ovšem v plné míře simulována a ověřena tato možnost.

---

na vytvoření webového serveru pro strojové učení pomocí frameworku `flask`, nebo `https://www.djangoproject.com/start/` – tutoriál k frameworku `django`.

4.2.5 Databázový model

Data na serveru byla uložena ve formě SQLite databáze, která je nativně podporována v rámci `django` frameworku. Aby bylo možné data do databáze uložit, je potřeba vytvořit příslušné databázové modely. Například data z aktigrafu získaná za jeden časový okamžik a popsaná tabulkou 3.1 by byla modelována pomocí třídy jazyka Python uvedené ve výpisu 4.1.[4]

```
from django.db import models


class DataEntry(models.Model):
    time_stamp = models.DateTimeField('timestamp (yyyy-MM-dd HH:mm:ss:SSS)')
    x_axis = models.FloatField('x axis (g) (-8,8)')
    y_axis = models.FloatField('y axis (g) (-8,8)')
    z_axis = models.FloatField('z axis (g) (-8,8)')
    light_level = models.FloatField('light level (lux) (0,5000)')
    button_state = models.BooleanField('button state (0 -> false -> untoggle or 1 -> true -> toggle)')
    temperature = models.FloatField('temperature (deg C) (0,70)')
```

Výpis 4.1: Databázový model dat z aktigrafu

V rámci reálné implementace jsou data uložena na serveru ve formě CSV souborů, není třeba ukládat do databáze jednotlivé časové záznamy. Je ovšem potřeba vytvořit model pro celý datový CSV soubor dle výpisu 4.2.

```
from django.db import models
from django.core.validators import FileExtensionValidator


class CsvData(models.Model):
    subject = models.ForeignKey(Subject, on_delete=models.CASCADE)
    BODY_LOCATIONS = [
        ('L', 'Left wrist'),
        ('R', 'Right wrist'),
        ('O', 'Other')
    ]
    body_location = models.CharField('body location', max_length=1, choices=BODY_LOCATIONS, default='L')
    data = models.FileField('data', upload_to='data/', validators=[FileExtensionValidator(["csv"])])
    description = models.CharField('description', max_length=255, blank=True)
```

[4]Styl formátování kódu byl přebrán ze stránky `https://www.overleaf.com/project/5dc6c8e7ff0cbd0001062926` a bude použit i v ostatních výpisech.

```
15      creation_date = models.DateField('date of upload', auto_now_add
    =True)
16      graph_created = models.BooleanField('graph image created',
    editable=False, default=0)
17
18      def __str__(self):
19          return f'CSV data from subject {self.subject.code}  {self.
    description}'
```

Výpis 4.2: Databázový model datového souboru

Dále pak definujeme model pro subjekt, kterému data patří, a podobně pro další objekty, které chceme uchovat v databázi.

4.2.6 Architektura webové stránky

Webovou stránku vytvořenou v rámci diplomové práce můžeme rozdělit na tři logické celky: domovskou stránku, dashboard a administrativní stránky.

Domovská stránka představuje uživateli samotný projekt, je zde popsána funkcionalita a jsou uvedeny odkazy na další zdroje informací. Také jsou zde, tak jako na všech podstránkách, uvedeny citace na podpůrné materiály (ikony, šablona...). Těmto citacím, které jsou potřeba pro dodržení licenčních smluv, se bude věnovat mimo jiné kapitola 7. Domovská stránka je zobrazena na obr. B.1 v přílohách.

Dashboard přináší přehled výsledků jednotlivých subjektů. Data zde nelze upravovat ale pouze prohlížet. Dashboard zobrazuje v první řadě přehled všech subjektů, kde je ke konkrétním osobám uveden vždy jen unikátní identifikační kód. Tato stránka je přístupná komukoliv. Po kliknutí na konkrétní subjekt je zobrazena stránka s podrobnějšími údaji. Na této stránce lze nalézt věk subjektu, unikátní identifikační kód, zhodnocení poruch spánku a případnou diagnózu či další poznámky. Dále jsou zde zobrazena zpracovaná data o spánku z aktigrafu, a to formou grafů. Jelikož tato stránka s detaily obsahuje potencionálně zneužitelné informace, je zabezpečena. Zabezpečení je pak věnována vlastní kapitola 4.2.10. Datům samotným z pohledu ochrany osobních údajů se bude dále věnovat kapitola 7.4. Na obr. B.2 a obr. B.3 v přílohách je zobrazena část stránky Dashboard, konkrétně přehled subjektů a detail jednoho zvoleného subjektu.

Posledním logickým celkem je administrace. Tato část umožňuje přidávat, upravovat a mazat subjekty a přiřazovat jim data. Stránka je klíčová co se týče bezpečnosti a bude jí věnována kapitola 4.2.8.

Dále pak stránka obsahuje ještě čtvrtý logický celek, stránku věnovanou zpracování dat pomocí strojového učení. Ta je ovšem ve výchozím stavu skrytá a nedostupná, zobrazena a zpřístupněna je pouze přihlášeným privilegovaným uživatelům systému. Této stránce se věnuje kapitola 5.2.

Všechny webové stránky byly vytvořeny pomocí responzivního web designu, tedy jsou renderovány odlišným způsobem v závislosti na velikosti display. To umožňuje vytvořit vhodné zobrazení jak na monitoru notebooku, kdy je kupříkladu zobrazeno 6 sloupců subjektů na stránce Dashboard, a zároveň i na chytrém telefonu, kdy jsou zobrazeny sloupce pouze dva. Rovněž velikost textu či obrázků je přizpůsobována dynamicky velikosti displeje.

4.2.7 Zabezpečení, uživatelé a skupiny

Pro zajištění bezpečnosti je na serveru vedena databáze uživatelských účtů. Pro autentizaci je potřeba zadat správné přihlašovací jméno a heslo. Po úspěšné autentizaci je uživateli vygenerováno `sessionId` a uloženo `cookies` do jeho webového prohlížeče. Není tedy třeba provádět přihlášení pro každou operaci. Implementace je řešena nativně frameworkem `django`, bezpečnost je ověřena a otestována řadou vývojářů a uživatelů.

Heslo je v databázi uloženo ve formě hashe. Je použita hashovací funkce SHA-256 a technika „solení hashe“. Na výpise 4.3 je demonstrováno uložení hesla s reálnými daty, a je rovněž demonstrována aktualizace hashe po změně hesla.

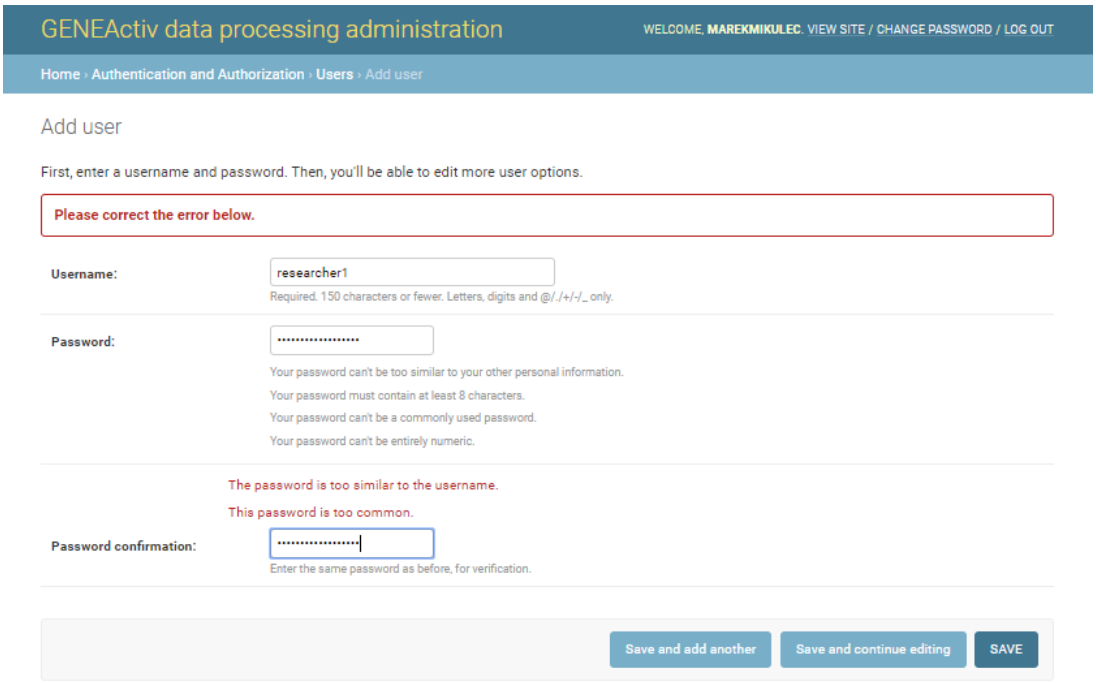
```
>>> from django.contrib.auth.models import User
>>> print(User.objects.all().first().password)
pbkdf2_sha256\
$150000$3iDcP279a1iU$dsGzvwl9k3NL4yrJl0jk5WeCpVlhebvrsmXd5LxZJ70=
>>> # Password changed via GUI
>>> print(User.objects.all().first().password)
pbkdf2_sha256\
$150000$OGseeoYHwS6L$G9jlHkdu7JxFwpEAoLca70KGwclFUKw1NUae6tLhEjY=
```

Výpis 4.3: Forma uložení hesel v databázi

Při tvorbě hesla jsou aplikovány validace bezpečnosti hesla, které lze dále nastavit a ještě více zpřísnit. Praktická ukázka je zobrazena na obr. 4.1, kde byla snaha nastavit uživateli `researcher1` heslo `Researcher`. Uživatel tak nemůže jednoduše kompromitovat bezpečnost tím, že by si přenastavil heslo vlastním slabým heslem, je donucen dodržet bezpečnostní politiku systému.

Rovněž byly vytvořeny skupiny uživatelů a těmto skupinám byla přiřazena různá práva (možné je také přiřadit práva přímo konkrétnímu uživateli). Dle tab. 2.1, kde byly definovány povinnosti a zodpovědnosti jednotlivých procesních rolí, byly vytvořeny tři skupiny uživatelů.

První je skupina administrátorská, kam spadá role správce serveru a role koordinátora studie. Uživatelé patřící do této skupiny mají plná práva pracovat s daty a rovněž spravovat uživatelské účty a skupiny samotné. Mají také právo nahlížet do dat všech subjektů.



Obr. 4.1: Validace bezpečnostních politik hesla

Druhá skupina je skupina výzkumných pracovníků. To jsou pověřené kontaktní osoby, které shromažďují data a mohou je nahrávat na server. Mají tedy možnost spravovat subjekty studie a jejich data. Tato skupina ovšem nemá práva spravovat uživatelské účty a skupiny. Rovněž nemají možnost prohlížet výsledky studie jednotlivých subjektů, viz podkapitola 4.2.10.

Třetí skupina, subjekty studie, pak má možnost navštívit stránku dashboardu, kde se dozví své výsledky díky znalosti specifického kódu subjektu, který jim patří, a znalosti svého hesla. Subjekty v rámci administrace webové stránky nemají žádné pravomoci. Rovněž jim není povoleno prohlížet data ostatních subjektů. Tím jsou naplněny požadavky dle RACI matice 2.1.

V systému je tedy při správném fungování zajištěna dodatečná bezpečnost tím, že jsou vhodně rozděleny pravomoci. Administrátor systému může provádět administrativní úkony a zpracovávat data. Zároveň ovšem o samotných subjektech nemá informace, na základě kterých by je mohl identifikovat, neboť jsou data pseudonymizována. Informacemi o identitě subjektu naopak disponuje výzkumný pracovník, typicky ošetřující lékař. Ten je schopný uživateli v systému vytvořit psedonymizovaný účet, vybavit jej aktigrafem a zajistit data. Není ovšem schopný nahlížet do dat subjektů která byla zpracována pomocí strojového učení (po změně hesla uživatelem). Ta jsou dostupná pouze samotnému uživateli, kterému přísluší, či případně administrátorovi systému, který k datům přistupuje z důvodu administrace

či vědeckého výzkumu.

V rámci budoucí práce lze bezpečnost dále zvýšit dle zdroje [12].

4.2.8 Zabezpečení stránky administrace

Pro přístup do Administrace je potřeba se autentizovat uživatelským jménem a heslem, jak bylo popsáno výše. Následně je zde možnost upravovat nastavení dle nastavených práv a spravovat datovou základnu studie. Uživatelé, skupiny, uživatelská práva, subjekty, vše bylo v podstatě vytvořeno skrze administraci. Jedná se tedy o klíčovou funkcionalitu celé webové stránky.

Celá stránka je vytvořena standardizovaně skrze framework `django`. Díky tomu je zajištěna vysoká míra stability a bezpečnosti. `Django` zajišťuje například ochranu proti *Cross Site Scripting – vložení skriptu z cizího zdroje, který je následně vykonán prohlížečem na straně uživatele* (CSS), proti *Cross Site Request Forgery – zneužití uživatelských oprávnění autentizovaného uživatele k vykonání nechtěných akcí bez vědomí uživatele* (CSRF) a dalším [13].

4.2.9 Zabezpečený přenos souboru

Data subjektů lze nahrát přes stránku administrace, která je dostupná pouze po přihlášení. Tak je zajištěno, že data budou nahrávat pouze oprávněné osoby. Data jsou v databázi spojena se subjektem, kterému patří. V databázi ovšem není fyzicky uložen soubor, pouze cesta k němu. Soubor se fyzicky zapisuje do složky `data`. K tomuto souboru se následně přistupuje při dalším zpracování. Práva k souboru má pouze majitel procesu, pod kterým je spuštěný webový server, což dále snižuje riziko zneužití. Při nahrávání se validuje typ, povoleny jsou pouze soubory s příponou csv. Validátor frameworku `django` by měl zabránit nahrání jiných potencionálně škodlivých souborů na server.

4.2.10 Zabezpečení stránky detaily subjektu

Stránky s detaily subjektů byly z důvodu zachování anonymity zabezpečeny tak, aby byly přístupné pouze subjektům a administrátorům. Subjektům je zřízen účet se stejným uživatelským jménem, jako je kód subjektu, a bezpečným způsobem jim musí být předáno heslo. Po kliknutí na detail subjektu je uživatel přesměrován na přihlašovací stránku a posléze je mu přístup povolen či zamítnut. Účet může vytvořit pouze uživatel ze skupiny administrátorů. To jsou také jediní další uživatelé, kteří mohou stránky navštívit, aby ověřili data zde uvedená a zajistili případné opravy. Taktéž jsou tyto události logovány, tedy lze zpětně dohledat, kdo s jakým uživatelským jménem si kdy prohlížel která uživatelská data.

Původně byl v rámci implementace zpracování dat pomocí strojového učení generován graf, který znázorňoval informace o spánku subjektu, uložen jako obrázek. Z tohoto obrázku se poté graf načítal na webovou stránku detailu subjektu. Toto řešení ovšem představovalo bezpečnostní riziko, neboť se znalostí příslušné URL adresy bylo možné k obrázku přistoupit přímo, a tím získat potencionálně citlivé informace. Z toho důvodu byla následně použita jiná varianta. Grafy jsou generovány dynamicky až po dotazu na příslušnou stránku, nejsou nikde uloženy, aby je nebylo možné zcizit, a jsou zobrazeny pomocí interaktivního doplňku knihovny `plotly`. Graf je zobrazen pomocí `javascriptu`, umožňuje provádět interaktivní podrobnou analýzu.

4.2.11 Logování

V rámci zajištění bezpečnosti je potřebné zajistit dostatečné logování událostí. Logování je také užitečné pro hledání chyb a podobně. Proto byla logování věnována patřičná pozornost.

Samotná úroveň logování závisí od toho, zda se nacházíme v globálním `django` nastavení `debug`. Pokud je režim `debug` zapnutý, je logováno více informací, které slouží především pro vývojáře. V reálném prostředí ovšem musí být parametr `debug` nastaven na nepravdu a následně probíhá pouze logování patřičných událostí.

Logování používá tři výstupy. Logování do konzole, logování do souboru a v případě chyby či výjimky odeslání emailu na email administrátora. Tak je zajištěna potřebná dokumentace o činnosti systému a je zajištěna co nejrychlejší reakce ze strany administrátora v případě nefunkčnosti či chyby.

4.2.12 Automatické testování

Automatické testování probíhalo formou unit testů. Zajímavostí v případě frameworku `django` je i fakt, že lze testovat třídy `view`, tedy lze pomocí unit testů prověřit funkcionalitu tak, jak ji vidí koncový uživatel, a zároveň je pomocí klasických unit testů možno ověřit kvalitu kódu.

Implementace obsahuje automatické testy především pro část zpracování dat. Je testována konzistence a správný formát dat. Formou TDD – Test Driven Development – pak probíhala implementace některých částí systému.

Automatické testy pro konzistenci dat byly zpřístupněny přes UI, neboť jsou extrémně užitečné v procesu zpracování dat, viz kap. 5.2.

4.2.13 Zdrojové kódy

Zdrojové kódy projektu lze nalézt na následujícím odkaze: `https://gitlab.com/MarekMikulec/geneactiv-processing-data`. Na odkaze naleznete poslední stabilní verzi, která je obsažena ve vývojové větvi `master`. Pro možnost prezentovat zdrojový kód v podobě, v jaké se nacházel v době odevzdání diplomové práce, byla vytvořena vývojová větev `diploma-thesis` dostupná na odkaze `https://gitlab.com/MarekMikulec/geneactiv-processing-data/tree/diploma-thesis`. Zde lze ověřit, že poslední změny pochází z doby před termínem odevzdání práce.

4.3 Klientská část systému

V rámci klientské části z pohledu výzkumného pracovníka není třeba další implementace. Jak bylo řečeno v rámci podkapitoly 4.1, je vyžadován jakýkoliv počítač, který je kompatibilní s GENEActivPcSoftware, pro stažení dat z aktigrafu, tedy v podstatě libovolný počítač s rozhraním USB a operačním systémem Windows verze 7 a výše. Dále je pak vyžadován standardní webový prohlížeč kompatibilní se značkovacím jazykem HTML5. Potřebná je pouze znalost příslušné webové stránky a přihlašovacích údajů.

Z pohledu klienta jako subjektu studie je vyžadováno jakékoliv zařízení s webovým prohlížečem kompatibilní se značkovacím jazykem HTML5. Využít lze nejen počítač, ale i chytrý telefon či tablet. Není třeba žádné instalace, stačí vstoupit na příslušné webové stránky a mít potřebnou znalost přihlašovacích údajů. Stránky jsou optimalizovány pro použití na rozdílných velikostech displeje, proto budou vhodně zobrazeny na počítači i na mobilním telefonu.

5 Implementace zpracování dat a evaluace klasifikačního modelu

V rámci zpracování dat si práce klade za cíl stanovit čas spánku z dat zaznamenaných pomocí aktigrafu. Implementace zpracování dat bude popsána v jednotlivých krocích, tak jak je definována výčtem v kapitole 3.3.2. Zvolená metoda a model budou následně blíže specifikovány dle rozdělení uvedeného v kapitole 3.3.1.

5.1 Cíl: stanovení času spánku

Cílem implementace je stanovit čas spánku subjektu pomocí dat z aktigrafu.

Data obsažená v datasetu [14] obsahují vždy časovou řadu záznamů z aktigrafu, odpovídající jedné noci spánku subjektů. Dále dataset obsahuje data získaná téže noci pomocí polysomnografie [14, 15].[1] Data polysomnografie obsahují informaci o spánku subjektů a budou použita jako referenční. S pomocí těchto dat bude model natrénován, aby byl schopný rozpoznat čas spánku subjektů.

5.2 Uživatelské rozhraní

Pro správu studie byla vytvořena webová stránka utilit, která umožňuje spravovat veškeré akce zpracování dat. Stránka je dostupná pouze s administrátorským účtem. Pro ostatní uživatele je stránka skryta, a rovněž snaha přistoupit na stránku pomocí URL je zamítnuta bez přihlášení k administrátorskému účtu. Tyto restrikce jsou aplikovány z důvodu bezpečnosti a stability, jelikož prováděné operace jsou určeny pouze poučeným uživatelům systému.

Na stránce lze spustit operace předzpracování dat, parametrizaci dat, tréning modelu strojového učení a vytvoření predikcí pro všechna data na serveru. Každá z uvedených operací může být provedena zvlášť, či mohou být všechny provedeny naráz. Jednotlivé operace budou popsány v následujících kapitolách.

O výsledcích je uživatel informován webovým rozhraním, podrobné informace jsou zapisovány do logu.[2] Obr. B.4 v přílohách ilustruje vzhled webové stránky utilit.

Jak je z obr. B.4 patrné, mimo samotné operace předzpracování, parametrizace, trénování modelu a vytvoření predikcí byla ke každé z těchto operací ještě doplněna

[1] Polysomnografie je metoda pro vyšetření spánku, která využívá současného snímání více fyziologických parametrů pacienta. Často se využívá speciálních přístrojů, například elektroencefalogram (EEG), elektrokardiogram (EKG) a podobně [16].

[2] Log který vznikl při postupné práci na systému je dostupný na webové stránce `https://gitlab.com/MarekMikulec/geneactiv-processing-data/-/blob/master/www/debug.log`. Log obsahuje spoustu důležitých informací a může sloužit jako dodatečná dokumentace systému.

kontrolní funkce a funkce pro smazaní dat. Kontrolní funkce kontroluje přítomnost uložených výsledků dané operace, správnost a konzistenci těchto výsledků. Uživatel tak může zjistit, které kroky již byly provedeny, a zda jsou výsledky v pořádku a data nebyla ze serveru například smazána neodborným zásahem. Funkce mazání umožňuje smazat všechny uložené výsledky. Bez smazání předchozích výsledků jsou při opakovaném spuštění operací zpracována pouze nová data, proto je pro kompletní opakování data třeba smazat. Obě tyto funkce významnou měrou usnadňují práci se systémem a byly velmi užitečné i při samotném vývoji.

5.3 Čištění dat

Čištění dat probíhalo v rámci implementace až v kroku předzpracování dat, kde byla předzpracována pouze relevantní data. Zároveň byly algoritmy automatického předzpracování odhaleny chyby v datech, například chybné datum měření v souboru s polysomnografickými daty subjektu MECSLEEP29. Chyby bylo možné opravit, data nemusela být z trénovací množiny vyřazena. V rámci dat obsažených v datasetu tak nebylo třeba odstraňovat nekonzistentní data. Data byla ve své plné podobě připravena k integraci.

5.4 Integrace dat

Integrace dat byla provedena za pomoci serverové implementace. Data byla na server nahrána a integrována stejným způsobem, jako by to provedl výzkumný pracovník, viz definice rolí dle tab. 2.1.

5.4.1 Data aktigrafu

Data z aktigrafu GENEActiv byla převedena na jednotný formát pomocí nástroje `csv converter`, obsaženého v rámci aplikace GENEActivPCSoftware. Pomocí automatického testu bylo ověřeno, že vzorkovací frekvence ve všech souborech odpovídá hodnotě 85Hz.

Podoba dat po vynechání úvodní informační hlavičky souboru je uvedena v následujícím výpise 5.1.

```
2013-06-11 02:50:58:505,0.9573,0.2857,0.0989,0,0,31.4
2013-06-11 02:50:58:516,0.9573,0.2779,0.1067,0,0,31.4
2013-06-11 02:50:58:528,0.9612,0.2741,0.1107,0,0,31.4
2013-06-11 02:50:58:540,0.9651,0.2974,0.1185,0,0,31.4
2013-06-11 02:50:58:551,0.9495,0.2779,0.1107,0,0,31.4
```

Výpis 5.1: Ukázka části csv souboru záznamů z aktigrafu

Formát dat je po integraci následující:

- časové razítko ve formátu %Y-%m-%d %H:%M:%S:%L,[3]
- zrychlení v ose X,
- zrychlení v ose Y,
- zrychlení v ose Z,
- úroveň osvětlení,
- stav tlačítka,
- teplota.

Formát odpovídá teoretickým parametrům aktigrafu uvedeným v tab. 3.1.

5.4.2 Data polysomnografie

Data z polysomnografie obsažená v rámci datasetu mají všechna stejný formát. Jsou uložena jako textové soubory. Strukturou v podstatě odpovídají souborům CSV, pouze místo čárky zde k oddělení hodnot slouží znak tabulátoru.

Ukázka dat po vynechání informační hlavičky představuje následujícím výpise 5.2.

```
1     N3   Right  23:05:40   N3   30
2     N3   Right  23:06:10   N3   30
3     N3   Right  23:06:40   N3   30
4     N1   Right  23:07:10   N1   30
5     W Right  23:07:40   W 30
6     W Right  23:08:10   W 30
7     W Supine   23:08:40   W 30
8     W Supine   23:09:10   W 30
9     W Supine   23:09:40   W 30
10    W Supine   23:10:10   W 30
```

Výpis 5.2: Ukázka části textového souboru záznamů z polysomnografie

Význam dat je následující:

- fáze spánku dle klasifikace aktigrafie:
  - W = Wake – stav bdělosti,
  - N1 = Non-REM 1 – fáze přechodu mezi bděním a spánkem,
  - N2 = Non-REM 2 – hlavní fáze spánku,
  - N3 = Non-REM 3 – fáze hlubokého spánku,
  - R = REM – fáze spánku, při které se zdají sny [17, s. 194],
- poloha subjektu,
- časové razítko ve formátu %H:%M:%S,
- událost = fáze spánku dle klasifikace aktigrafie,

[3]Formát času lze ověřit například na stránce `http://www.strfti.me`.

- délka časového okna v sekundách.

V rámci integrace dat bylo potřeba provést opravu v souboru polysomnografie subjektu MECSLEEP34, jelikož v datech nebyla uvedena informace o poloze subjektu. Za účelem sjednocení formátu byla přidána zástupná hodnota a adekvátním způsobem byla upravena i hlavička souboru.

5.5 Výběr dat

Z pohledu datasetu budou využity všechny dostupné datové soubory. Jedná se tedy o 28 subjektů, 55 souborů z aktigrafu a 28 souborů polysomnografie. Všechna data obsažená v souborech aktigrafu a polysomnografie ovšem nejsou pro zpracování relevantní.

5.5.1 Data aktigrafu

V rámci dat z aktigrafu byla vybrána následující data:

- časové razítko ve formátu %Y-%m-%d %H:%M:%S:%L,
- zrychlení v ose X,
- zrychlení v ose Y,
- zrychlení v ose Z,
- teplota.

Další veličiny budou při předzpracování zanedbány. Jedná se o stav tlačítka, kdy tento údaj nemá pro studii žádnou vypovídací hodnotu, a o úroveň osvětlení, neboť trénovací data byla získaná v laboratorních podmínkách, a neodpovídají hodnotám reálného prostředí. Obě hodnoty jsou téměř vždy nulové.

5.5.2 Data polysomnografie

V rámci dat z polysomnografie, která budou použita jako referenční výsledky, byla vybraná tato data:

- fáze spánku dle klasifikace aktigrafie (W, N1, N2, N3, R),
- časové razítko ve formátu %H:%M:%S, doplněné o datum %d/%m/%Y obsažené v hlavičce souboru.

Jelikož je cílem stanovit dobu spánku subjektu, budou dále fáze spánku N1, N2, N3 a R chápány jako spánek a fáze spánku W jako stav bdělosti. U časového razítka bylo rovněž provedeno navýšení data při přechodu skrze půlnoc. Zanedbána byla poloha subjektu, neboť není pro studii relevantní; událost, protože nese stejnou hodnotu jako fáze spánku; a dále pak délka časového okna, jelikož se jedná o neměnnou konstantu 30 s.

5.5.3 Stanovení společného časového okna

Počátek a konec záznamu dat z aktigrafu a dat polysomnografie není shodný, někdy i v řádu hodin. Proto jsou automaticky vybrána pouze data z aktigrafu a data z polysomnografie, která byla měřena současně. Ostatní data nebudou dále použita, neboť by chyběl buď výsledek v podobě dat polysomnografie, nebo by chyběla naopak data pro zpracování, která mají poskytnou výsledek shodný s výsledkem polysomnografie.

5.5.4 Snížení počtu vzorků dat aktigrafie

Zdroj [18] se zabývá analýzou velkého množství vědeckých prací, které zpracovávají data z aktigrafů metodami strojového učení. Tento zdroj uvádí, s odkazem na analyzované studie, že dostatečná vzorkovací frekvence aktigrafu se pohybuje kolem 25 Hz. Vyšší vzorkovací frekvence údajně nepřináší žádný benefit. Z toho důvodu může být vzorkovací frekvence dat snížena z původních 85 Hz na třetinu, tedy na přibližně $28,5$ Hz. Tím lze docílit výrazné úspory paměti a snížení výpočetní náročnosti. Čas potřebný pro extrakci „features“ – parametry dat, se díky snížení počtu vzorků sníží rovněž o třetinu. Předzpracování dat a parametrizace pro data za 55 nocí tedy na jednom jádru procesoru s frekvencí $2,7$ GHz trvá místo přibližně tří hodin pouze jednu hodinu.

5.6 Předzpracování dat

Před samotnou aplikací učícího algoritmu je nutné data předzpracovat. Prvním důvodem je nevýhoda načítaní dat přímo ze souborů csv a txt. Jde o pomalejší operaci, které je lepší se vyhnout. Dalším důvodem jsou redundantní data. V rámci předzpracování se z dat extrahují parametry a algoritmus strojového učení dále pracuje pouze s těmito parametry.

5.6.1 Předzpracování dat z textových souborů

Úplně prvním krokem předzpracování dat je oprava dat předzpracovaných pomocí aplikace GENEActivPCSoftware. Ta totiž v případě nevyplnění některého z nepovinných údajů dosazuje na dané místo v CSV souboru tzv. `NULL` symbol `\x00`. Jelikož je `NULL` symbol použit k ukončení řádku, je takto vytvořený soubor považován za poškozený. Proto je třeba nejprve všechny symboly `\x00` nahradit mezerou. Po této opravě již mohou být data z CSV souboru bez problému načtena.

Při samotném předzpracování jsou zohledněna pouze vybraná data vyjmenovaná v předcházející kapitole. Rovněž je snížena vzorkovací frekvence na třetinu, zohledněný je tak každý třetí datový záznam. Data jsou převedena z textových řetězců do struktur jazyka `python`. Časová razítka jsou převedena s použitím zmíněných formátů do objektu `datetime`, ostatní hodnoty jsou převedeny na číselný typ `float` či `int`. Celý soubor je tak v paměti reprezentován listem jazyka python s daty adekvátních datových typů.

Takto připravený list je uložen na disk pomocí knihovny `pickle`. Knihovna slouží k serializaci a deserializaci objektů jazyka `python`. „Cachováním“ dat se ušetří množství času, jelikož při dalším dotazu na data je rovnou použit předzpracovaný soubor, pokud existuje.[4] Zároveň je možné smazat původní data a ponechat pouze serializované objekty. Potřebná pamět je 10× menší.[5]

Pro další detaily předzpracování lze nahlédnout do zdrojového kódu na odkaze `https://gitlab.com/MarekMikulec/geneactiv-processing-data/-/tree/master/www/mysite/dashboard/logic/preprocessing` .

5.6.2 Parametrizace dat

Pro efektivní aplikaci algoritmů strojového učení se data statisticky předzpracovávají. Statistické parametry dat – features – jsou finální vstup algoritmu strojového učení. V rámci předzpracování jsou stanoveny statistické parametry pro každý 30vteřinový úsek dat. Zvlášť jsou zpracována data z tří os akcelerometru a zvlášť jsou vyhodnocena data z teplotního čidla. V případě dat akcelerometru se z hodnot zrychlení v ose x, y a z spočte velikost vektoru, magnitude, podle vzorce

$$Magnitude = \sqrt{x^2 + y^2 + z^2}.$$

Dále jsou pak extrahovány statistické parametry pro vektor velikostí vektorů za jednotlivé časové okamžiky ($n = 857$). Vektor záznamů dat o teplotě je statisticky zpracován přímo ($n = 857$). Pro oba tyto vektory – statistické soubory – je stanoveno 45 statistických parametrů. Parametry jsou vyjmenovány v tab. C.1 v přílohách.

Data pro statistické zpracování jsou načtena z předzpracovaných souborů, viz kap. 5.6.1.

K výpočtu parametrů byly využity knihovny `statistics`, `scipy` a `numpy`. Spočtené statistické parametry pro každý soubor jsou shromážděny v objektu `DataFrame` knihovny `pandas`. Tak je možné s daty dále pohodlně pracovat. Objekt `DataFrame` lze jednoduše uložit a načíst do souboru `.xlsx` aplikace Microsoft Excel. Vzniká

[4] Předzpracování dat z textových souborů pro celý dataset 55 nocí na jednom jádru procesoru s frekvencí 2, 7 GHz trvalo 35 minut. Deserializace objektů se pohybuje v řádu sekund.

[5] Původní datové soubory dosahují velikosti 20 GB, serializované objekty dosahují velikosti 2 GB.

rozsáhlá tabulka s časovými razítky a jednotlivými parametry. Ke každému souboru z aktigrafu je tedy vytvořena tabulka parametrů, která je opět uložena na serveru pro zrychlení dalších výpočtů. Následně jsou pak všechny parametry ze všech souborů spojeny do jednoho `DataFrame`, který je finální množinou vstupů pro algoritmus strojového učení. Informace o datasetu získané metodou `info` lze nalézt ve výpise C.1 v přílohách práce.

Finální dataset je uložen na serveru jako `.xlsx` soubor, opět kvůli zrychlení následných operací. Načtení dat z tabulkového souboru oproti opětovné parametrizaci přináší 50násobné zrychlení.[6] Teoreticky je rovněž možné smazat původní data a používat nadále pouze extrahované parametry. Z původních 20 GB čistých dat vzniká 37 MB velký dataset extrahovaných parametrů. Paměťové nároky tak 500krát klesly.[7]

5.7 Aplikace učících se algoritmů

Nejprve bude popsán princip, jakým způsobem by měl algoritmus pracovat a co je očekávaným výstupem. Následně bude dle kapitoly 3.3.1 klasifikován vhodný model. Poté bude zvolena konkrétní implementace.

5.7.1 Princip algoritmu

Vstupem pro algoritmus jsou předzpracovaná data z akcelerometru. Cílem je, aby algoritmus poskytl data odpovídající zjednodušeným datům polysomnografie. Pro každý 30vteřinový úsek tak algoritmus musí rozhodnout, zda subjekt spí, či nikoliv. Výstup pak bude porovnán s reálnými daty polysomnografie.

30vteřinové časové okno bylo zvoleno na základě dat polysomnografie, kdy záznamy mají 30vteřinové rozestupy. Dle zdroje [18] je u většiny obdobných studií časové okno voleno fixně, obvykle s velikostí 1–60 s. Volba tedy není v rozporu s osvědčenou praxí.

5.7.2 Zařazení algoritmu

Algoritmus bude zařazen na základě rozdělení popsaných v kapitole 3.3.1.

[6] Parametrizace 55 souborů na jednom jádru procesoru s frekvencí 2,7 GHz trvala přibližně 50 minut. Načtení dat z `.xlsx` souboru trvá přibližně minutu.

[7] Dataset je dostupný ke stažení na odkaze `https://gitlab.com/MarekMikulec/geneactiv-processing-data/-/blob/master/www/ml/dataset.xlsx` . Další podrobnosti o parametrizaci lze vyčíst ze samotné implementace na odkaze `https://gitlab.com/MarekMikulec/geneactiv-processing-data/-/tree/master/www/mysite/dashboard/logic/features_extraction` .

Co se týče míry lidského vstupu, algoritmus odpovídá charakteristice **strojového učení s učitelem**. Algoritmu jsou poskytnuty očekávané výsledky v podobě dat polysomnografie a je očekáváno adekvátní natrénování modelu.

Algoritmu jsou zasílána připravená data po dávkách. Nedochází k přeučování modelu na základě nově příchozích dat. Model bude natrénován a v té podobě zafixován do doby než bude spuštěno opětovné trénování nad celým dostupným datasetem. Jedná se tedy o **strojové učení po dávkách**.

Algoritmus je typu **učení založené na modelu**. Cílem je natrénovat model, který je u neznámých dat schopný rozpoznat, zda se jedná o data spánku či bdění.

5.7.3 Volba metody strojového učení

Při volbě modelu byly zohledněny výsledky průzkumu zdroje [18]. Ten cituje několik studií, které došly k shodným výsledkům s použitím různých metod strojového učení.

Následně jsou citovány další studie, které se od sebe výsledky velmi liší a každá preferuje jinou metodu. Závěr tedy není jednoznačný. Většina studií testovala více rozdílných metod. Otestovány byly například tyto metody strojového učení:

- *Artificial Neural Network – umělá neuronová síť* (ANN) (32 z 62 studií),
- *Support Vector Machines – metoda podpůrných vektorů* (SVM) (18 z 62 studií),
- *Random Forest – náhodný les* (RF) (12 z 62 studií),
- *Decision Tree – rozhodovací strom* (DT) (11 z 62 studií),
- *Linear regression – lineární regrese* (LR) (7 z 62 studií) [18].

Vzhledem k povaze úlohy byl zvolen algoritmus XGBoost, neboť tento algoritmus patří v současnosti k nejpoužívanějším klasifikačním metodám a dosahuje skvělých výsledků v porovnání s konkurencí.

5.7.4 XGBoost

XGBoost je metoda strojového učení založená na principu Tree boosting. Implementace je volně dostupná v podobě open source knihovny pro celou řadu programovacích jazyků včetně jazyka `Python`.

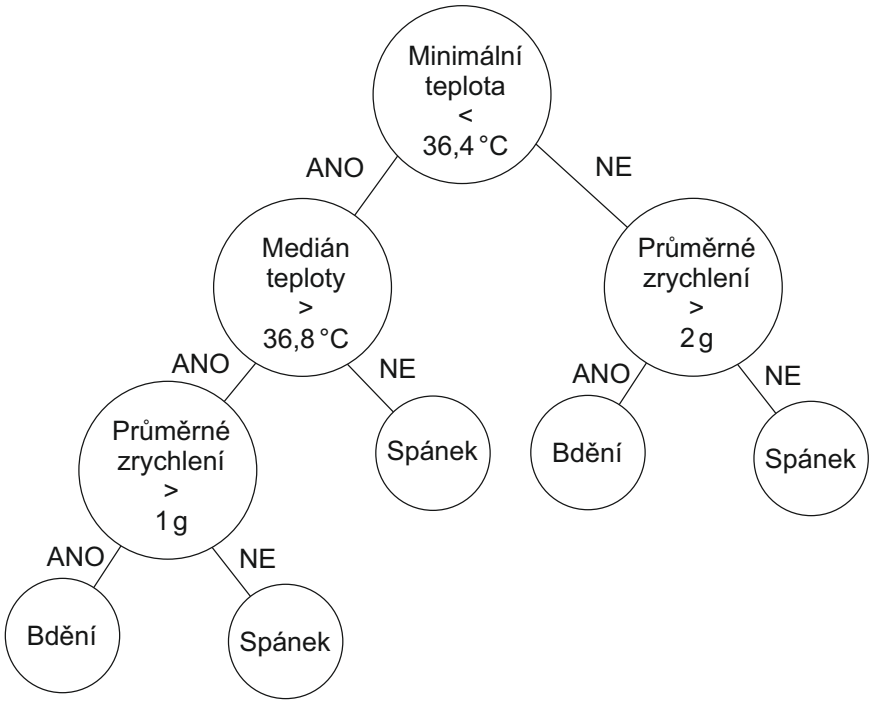
Metoda je hojně využívaná v celé řadě úspěšných studií. V prestižních soutěžích pořádaných například stránkou `https://www.kaggle.com`, se řešení využívající metodu XGBoost pravidelně umisťují na předních příčkách. Za úspěchem metody stojí v prvé řadě její velká škálovatelnost. Její použití je možné pro velkou řadu různých problémů. Díky optimalizované implementaci je metoda až desetkrát rychlejší než jiné oblíbené metody. Přináší možnost zpracovat velké množství dat na nespecializovaném hardware [19].

Metoda XGBoost byla zvolena díky zmíněným charakteristikám, dále pak kvůli absenci výkonného hardware pro implementaci jiných metod strojového učení, například neuronových sítí, díky relativně snadnému použití, podpory výpočtů s pomocí procesoru i s pomocí grafické karty, díky plné integraci s oblíbeným frameworkem `scikit-learn` a na doporučení vedoucího práce.

O implementaci metody XGBoost pojednává řada vědeckých publikací, především pak publikace [19], která má dle Google Scholar k 20. 5. 2020 5612 citací. Z toho důvodu není třeba implementaci opakovaně popisovat. Pro základní pochopení postačí, že algoritmus vytváří na základě vstupních dat rozhodovací stromy. Představíme li si například tři parametry:

- minimální teplota,
- průměrné zrychlení,
- medián teploty,

mohl by jednoduchý rozhodovací strom v našem případě vypadat například tak, jak je uvedeno na obr. 5.1.



Obr. 5.1: Rozhodovací strom

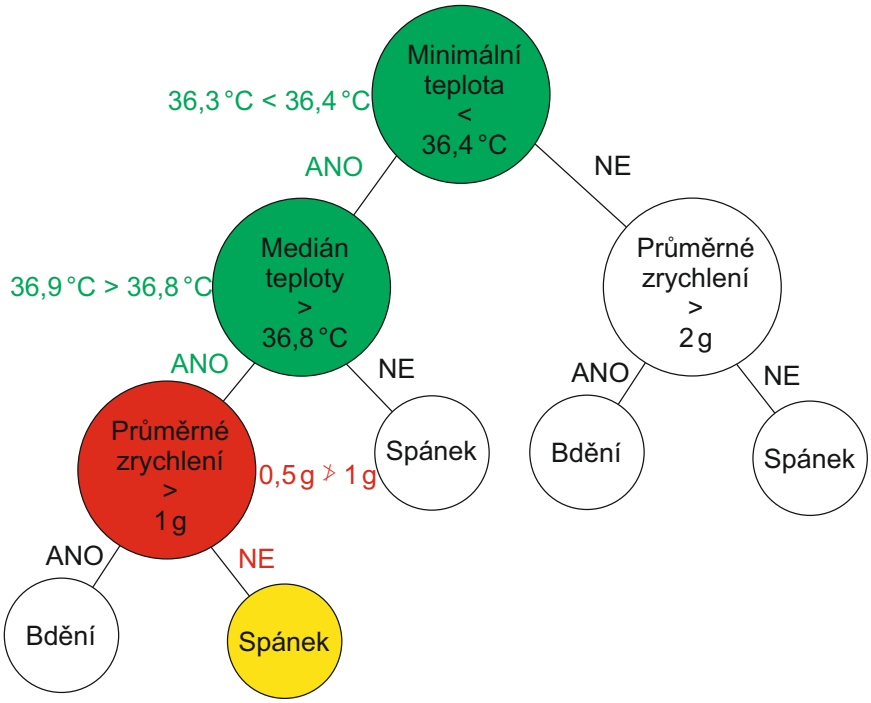
Mějme na vstupu rozhodovacího stromu 5.1 následující data:

Čas	Minimální teplota	Průměrné zrychlení	Medián teploty
22. 05. 2019 10:40:15	36,3 °C	0,5 g	36,9 °C

Pak je průchod stromem následující:

1. v prvním vrcholu (kořen stromu) je splněna podmínka: minimální teplota je nižší než 36,4 °C, postupujeme tedy hranou vlevo;
2. ve druhém vrcholu je splněn podmínka: medián teploty je vyšší než 36,8 °C, postupujeme tedy hranou vlevo;
3. ve třetím vrcholu není splněna podmínka: průměrné zrychlení je nižší než 1 g, postupujeme tedy hranou vpravo;
4. dostáváme se do koncového vrcholu stromu, vstupním datům přiřazujeme výsledek, jedná se o data spánku.

Pro lepší názornost je průchod zobrazen na obr. 5.2.



Obr. 5.2: Rozhodovací strom – příklad

Výstupem z příkladu by pak tedy byla informace, že v časový okamžik 22. 05. 2019 10:40:15 subjekt spal.

Algoritmus XGBoost využívá obdobné rozhodovací stromy.[8] V procesu učení modelu pak algoritmus stromy pozměňuje tak, aby dosáhl požadovaných výsledků

[8]Příklad stromu z modelu použitého v rámci systému lze nalézt na odkaze `https://gitlab.com/MarekMikulec/geneactiv-processing-data/-/blob/master/www/ml/trained/20-5-google-colab-logloss-cerror/decision_tree.pdf`. Rozhodovací strom je velmi rozsáhlý, především do šíře. Z toho důvodu není dokument vhodným způsobem zobrazen ve webovém rozhraní a je potřeba dokument stánout. Z důvodu rozsahu a poměru stran nebyl příklad rovněž umístěn do diplomové práce přímo a byl referován pouze skrze odkaz.

v rámci učení s učitelem. V rámci trénování modelu XGBoost využívá řadu optimalizačních technik například:

- **Shrinkage, Column Subsampling** – techniky k zabránění modelu k přílišnému natrénování na trénovací data (overfitting),
- **Approximate Algorithm** – technika pro rozdělení dat do skupin v případě že data nelze načíst celá do paměti, nebo v případě využití distribuovaného systému,
- **Weighted Quantile Sketch** – technika k efektivnímu rozdělení dat v rámci techniky Approximate Algorithm,
- **Sparsity-aware Split Finding** – technika zabraňující špatnému rozdělení dat z důvodu chybějících nebo nulových dat [19].

XGBoost je rovněž optimalizován, aby maximálně využil dostupný hardware. Dostupná je varianta implementace využívající grafické karty a varianta využívající pouze procesor. Obě varianty jsou implementovány vícevláknově, tedy využívají procesor na maximum. XGBoost rovněž optimalizuje práci s diskem. Zajímavá je rovněž možnost použít algoritmus XGBoost v distribuovaném systému, kdy je výpočetní zátěž rozložena mezi více zařízení [19].

5.7.5 Nastavení hyper-parametrů

Před samotným procesem trénování modelu XGBoost na vytvořeném datasetu je potřeba modelu nastavit tzv. hyper-parametry. Hyper-parametry nejsou modelem optimalizovány při procesu učení, jsou to parametry, které jsou nastaveny manuálně nebo automaticky před samotným započetím procesu učení. Nastavení hyper-parametrů může zásadním způsobem ovlivnit následný proces učení. Aby nebylo nutné hyper-parametry „tipovat“, či spoléhat na zkušenosti a znalosti experta, byly vyvinuty metody automatického nastavení hyper-parametrů [20].

Při implementaci byla použita knihovna XGBoost společně s knihovnou `scikit-learn`, která poskytuje implementované metody strojového učení. Pro hledání optimálních hyper-parametrů byla použita metoda `RandomizedSearchCV` z knihovny `scikit-learn`. Metoda umožňuje nalézt nejlepší hyper-parametry pomocí křížové validace (cross validation). Metoda umožnuje buď prozkoumat všechny dostupné hyper-parametry z určitého rozsahu, nebo je omezena počtem pokusů, nebo jsou stanoveny předem hodnoty hyper-parametrů, které mají být náhodně kombinovány.[9]

V rámci implementace systému byla využita poslední možnost. Výpis 5.3 je slovníkem testovaných hyper-parametrů, ze kterých má metoda `RandomizedSearchCV` vybrat nejlepší kombinaci.

[9]Dokumentaci metody `RandomizedSearchCV` lze nalézt na adrese `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html`.

```
param_grid = {
    "n_estimators": [200, 500, 1000],
    "learning_rate": [0.001, 0.01, 0.1, 0.2, 0.3],
    "gamma": [0, 0.10, 0.15, 0.25, 0.5],
    "max_depth": [6, 8, 10, 12, 15],
    "min_child_weight": [0.5, 1.0, 3.0, 5.0, 7.0, 10.0],
    "subsample": [0.5, 0.6, 0.7, 0.8, 0.9, 1.0],
    "colsample_bylevel": [0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0],
    "colsample_bytree": [0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0],
    "scale_pos_weight": [1, 3, 5, 6, 7, 9]
}
```

Výpis 5.3: Testované hyperparametry

Zbylé hyper-parametry byly nastaveveny staticky, dle cíle úlohy strojového učení, viz výpis 5.4.

```
model_params = {
    # General parameters
    "booster": "gbtree",
    "verbosity": 1,
    "n_jobs": -1,

    # Learning task parameters
    "objective": "binary:logistic",
    "eval_metric": ["error", "logloss", "auc"],
    "seed": 42,

    # Tree Booster parameters
    "n_estimators": 1000,

    # # GPU setting
    "gpu_id": 0,
    "tree_method": "hist",
}
```

Výpis 5.4: Statické hyper-parametry

Následuje výčet použitých hyper-parametrů s blyžším popisem dle [21].

- **General Parameters** – hlavní parametry
  - **booster** – volba boosteru, k dispozici jsou boostery `gbtree`, `gblinear` a `dart`; zvolen byl výchozí booster `gbtree`, který využívá standardní model založený na rozhodovacím stromě;
  - **verbosity** – volba podrobnosti logování, 2 = informační zprávy;
  - **n_jobs** – počet paralelních vláken, -1 = všechny dostupné dle hardware;
  - **n_estimators** – počet rozhodovacích stromů, které může model využít
  - **Parameters for Tree Booster** – parametry pro Tree Booster

        * **learning_rate** (eta) – upravuje rychlost, s jakou se model přizpůsobuje novým poznatkům; zabraňuje přeučení modelu, ale zpomaluje proces učení; nižší hodnoty umožňují rychlejší přizpůsobení modelu za cenu risku přeučení – overfitting;
        * **gamma** – obdobná funkce jako eta, složí k regulaci rychlosti přeučení modelu; čím je parametr vyšší, tím pomaleji se model přeučí na nové poznatky;
        * **max_depth** – maximální hloubka stromu (počet zanoření); vyšší hloubka může vést k lepší preciznosti, ale zároveň může způsobit přeučenní modelu;
        * **min_child_weight** – zjednodušeně řešeno minimální hodnota, které musí dosáhnout instance prvků v dané větvi stromu; pokud není hodnoty dosaženo, může se jednat o příliš specifickou podmínku, která přispívá k přílišnému přeučenní modelu;
        * **subsample** – výběr dílčího vzorku z dostupných dat, například při hodnotě 0,5 je použita polovina vzorků; zabraňuje přeučení;
        * **colsample_bylevel**, **colsample_bytree** – poměr úrovní/sloupců, který má být zvolen pro danou iteraci subsamplingu;
        * **tree_method** – metoda konstruování rozhodovacích stromů, na výběr jsou možnosti `auto`, `exact`, `approx`, `hist` a `gpu-hist`
            · `auto` – automaticky volí nejrychlejší metodu z následujících, pro malé datasety je zvolena metoda `exact`, pro větší datasety je zvolena metoda `approx`
            · `exact` – využívá všechny kandidáty,
            · `approx` – aproximační hladový algoritmus, využívá techniku Weighted Quantile Sketch a gradientní histogram, vhodné pro větší datasety,
            · `hist` – rychlejší varianta aproximačního hladového algoritmu,
            · `gpu_hist` – metoda `hist` akcelerovaná pomocí grafické karty;

          zvolen byl parametr `hist`, další informace jsou uvedeny v odstavci **Volba parametru `tree_method`**.
    - **scale_pos_weight** – parametr pomáhá zamezit přeučení na nevyvážených datasetech, například pokud dataset obsahuje 90% dat spánku a 10% dat bdělosti, může se model na takto nevyváženém datasetu naučit data v bdělém stavu úplně ignorovat; [22]
- **Learning Task Parameters** – parametry specifikující cíl učení
    - **objective** – cíl učení, k dispozici je celá řada možností viz [19]; zvolena hodnota `binary:logistic` = logistická regrese pro binární klasifikaci, více informací v odstavci **Binární logistická regrese**; dále byla také

otestována možnost `binary:hinge` = použita funkce hinge loss místo logistické regrese, funkce je orientovaná především na odhadování pravda/nepravda, než na určování pravděpodobností jako možnost předchozí, tato funkce se často používá s metodou Support Vector Machines – metoda podpůrných vektorů;

- **eval_metric** – evaluační metrika pro validaci výsledků, k dispozici celá řada možností; zvolena metrika `error` = výchozí metrika pro metodu `binary:logistic`; hodnota spočtena jako $\frac{\text{chybné případy}}{\text{všechny případy}}$; dále počítána i metrika `logloss` = záporná logaritmická cenová funkce (negative log-likehood),[10] a dále metrika `auc` = Area under Curve – oblast pod ROC křivkou, což je křivka, která vzniká, pokud vyneseme do grafu na jedné ose metriku **Specificity** a na druhé ose metriku **Sensitivity**;
- **seed** – výchozí hodnota pro funkci náhodných čísel

Volba parametru `tree_method`

Nejprve byl parametr `tree_method` nastaven na výchozí hodnotu `auto`. Ta volila automaticky metodu `exact`, která je vhodná pro malé datasety, jako ten využitý v rámci studie, je ovšem také výpočetně nejnáročnější metodou. Po přibližně 16 hodinách běhu metody `RandomizedSearchCV` byl pokus ukončen a bylo nutné přistoupit k optimalizacím, neboť proces hledání hyper-parametrů se ukázal jako časově velmi náročná operace. Aby algoritmus mohl vyhodnotit kvalitu jednotlivých parametrů, musí být jednotlivé modely s různými hyperparametry natrénovány a ohodnoceny [20].

Z důvodu urychlení procesu byla snaha využít grafickou kartu a zvolit jako parametr `tree_method` hodnotu `gpu_hist`, neboť zrychlení mělo být přiblyžně trojnásobné dle dokumentace dostupné na adrese `https://xgboost.readthedocs.io/en/latest/gpu/`. Jelikož v případě virtualizovaného operačního systému Fedora, na kterém byl celý systém vyvíjen, nelze skrze software Oracle Virtual Box využít externí grafickou kartu, muselo být celé řešení nainstalováno a překonfigurováno pro hostitelský operační systém Windows 10, který naštěstí podporuje nástroj Anaconda pro tvorbu virtuálních prostředí jazyka Python. Bylo třeba zkompilovat zdrojové kódy knihovny XGBoost pomocí programu Visual Studio. Dále bylo třeba nainstalovat výpočetní software CUDA a Intell Parallel Studio. Poté již bylo možné spustit algoritmus na grafické kartě.

Bohužel se ukázalo, že 2 GB paměti nejsou zdaleka dostačující, neboť je v rámci metody `RandomizedSearchCV` paralelně konstruováno množství modelů, a metoda

[10] Blyžší informace k funkci negative log-likehood jsou popsány v srozumitelné formě na odkaze `https://medium.com/deeplearningmadeeasy/negative-log-likelihood-6bd79b55d8b6`.

okamžitě končila s chybou přetečení paměti GPU. Celá snaha se tak ukázala jako zbytečná a akcelerace metody `RandomizedSearchCV` pomocí GPU jako neproveditelná.

Z toho důvodu byla nakonec zvolena hodnota `hist`, která využívá naplno procesor a pamět RAM a je optimálnější z hlediska výpočetní náročnosti než metoda `exact`. I přesto trval proces hledání hyper-parametrů při nastavení 500 rund na model v závěru 15 hodin při plném využití procesoru a paměti RAM. Parametry zařízení viz tab. A v přílohách práce.

Binární logistická regrese

Binární logistická regrese je funkce, která predikuje pravděpodobnost daného jevu striktně v rozmezí mezi 0 (jev nenastal) a 1 (jev nastal). Proto je vhodná pro binární klasifikaci. Výstup algoritmu je v našem případě interpretován následovně:

- 0 – jev nenastal = subjekt je vzhůru,
- 1 – jev nastal = subjekt spí.

Stejným způsobem jsou interpetovány i výsledky z algoritmu `logloss`.

Blíže o problematice logistické regrese pojednává například [23, s. 14–22] či [2, s. 181]. Implementační a matematické detaily zůstavají uživately knihovny XGBoost skrity, je tedy především potřeba ověřit na výsledcích správnost zvolené konfigurace.

Rozdělení datasetu – cross-validation

Pokud bychom použili stejná data pro trénování a testování, došli bychom do situace, kdy by model sice dosahoval perfektních výsledků na známých datech, ovšem ve chvíli, kdy by měl pracovat s daty neznámými, predikoval by model naprosto nesmyslné hodnoty. Tato situace se nazývá **overfitting** – přeučení, přetrénování.

Z toho důvodu je potřeba dataset vhodným způsobem rozdělit. Základní metodou, která bývá často použita, je rozdělení dat na data trénovací a data testovací. V použitém frameworku `scikit-learn` k tomu slouží například funkce `train_test_split`. Data jsou rozdělena procentuálně, například na 60% trénovací a 40% testovací data. Model je následně trénován na množině trénovacích dat a testován s pomocí dat testovacích. Tím je zabráněno jevu přeučení.

Pokud chceme testovat různé nastavení hyperparametrů, musíme opět ponechat část dat stranou, abychom na těchto neznámých datech mohli mezi sebou porovnat jednotlivé nastavení hyper-parametrů. Dataset je rozdělen na tři části:

1. trénovací data,
2. testovací data,
3. validační data.

V tuto chvíli jsme již schopni efektivně zhodnotit kvalitu nastavení jednotlivých hyper-parametrů [24].

Obrovskou nevýhodou je ovšem zmenšení jednotlivých setů, které můžeme použít. Dataset předpřipravený k učení modelu v rámci studie má 62 907 položek, viz výpis C.1 v přílohách práce. Jedná se tedy o poměrně malý dataset. Pokud bychom si představili hypotetický scénář rozdělení datasetu v poměru přibližně 60-20-20, tedy:

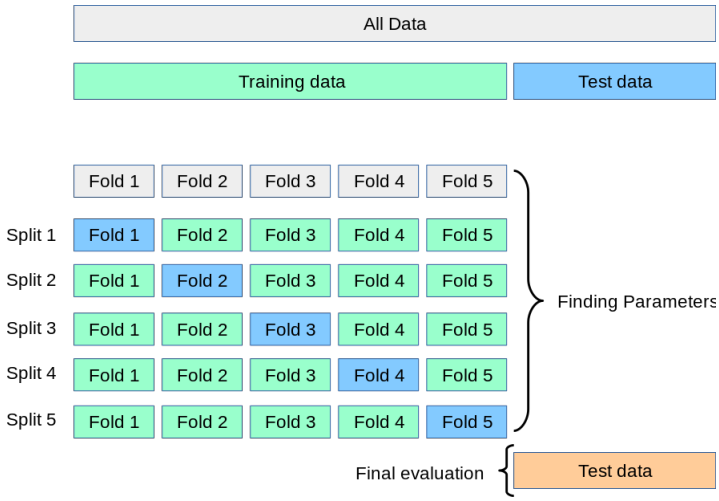
1. **trénovací data** – 60 % ≐ 37 745 datových záznamů,
2. **testovací data** – 20 % ≐ 12 581 datových záznamů,
3. **validační data** – 20 % ≐ 12 581 datových záznamů,

je zřejmé, že bychom pracovali s velmi malým množstvím dat. Řešením tohoto problému je metoda cross-validation (křížová validace).

V scénáři cross-validation rozdělíme dataset pouze na testovací a trénovací data. Validační data nejsou třeba, neboť místo nich využijeme křížovou validaci. Trénovací dataset rozdělíme na k malých dávek, k-folds, a provedeme tzv. k-násobnou křížovou validaci. Proces probíhá v k iteracích, v každé iteraci

- natrénujeme model na k - 1 dávkách dat,
- ověříme kvalitu natrénování modelu na dávce nechané stranou.

V další iteraci pak dávku ponechanou stranou použijeme jako trénovací a necháme stranou jinou dávku a tímto způsobem proces opakujeme právě v k iteracích, kdy každá dávka poslouží k validaci právě jednou. Výstupem procesu budou validační metriky vyjadřující kvalitu modelu při danném nastavení hyper parametrů.



Obr. 5.3: Křížová validace, zdroj [24]

Na závěr model natrénujeme na celém trénovacím datasetu a ověříme jeho kvalitu na testovacích datech. Testovací data byla v předchozích krocích ponechána stranou,

a jsou tak pro model neznámá. Tímto způsobem ověříme, že nedochází k přeučení modelu a principy byly modelem správně generalizovány. Celý proces je znázorněn na obr. 5.3 ze zdroje [24].

Metoda cross-validation byla využita při implementaci systému. Metodu není třeba implementovat, je součástí knihovny `scikit-learn`. K rozdělení do k dávek pro křížovou validaci slouží třída `RepeatedStatifiedKFold`, k provedení samotné křížové validace slouží metoda `cross_validate`.

5.8 Evaluace modelu strojového učení

5.8.1 Nalezené optimální hyper-parametry

Operace hledání hyper-parametrů byla provedena pro tři rozdílné konfigurace:

- **Model A**: `objective=binary:logistic`,
- **Model B**: `objective=binary:hinge`,
- **Model C**: `objective=binary:logistic` + dataset nadvzorkován do poměru 50 % spánek ku 50 % bdění, viz kap. 5.8.4.

Hyper-parametry s nejlepším výsledkem pro každý z modelů jsou uvedeny v tab. 5.1.

Tab. 5.1: Zvolené hyper-parametry

Parametr		**Zvolený parametr**		
Název	**Dostupné hodnoty**	**A**	**B**	**C**
n_estimators	[**100**, 200, 500, 1000]	1000	200	1000
learning_rate	[0.001, 0.01, 0.1, 0.2, **0.3**]	0.1	0.1	0.3
gamma	[**0**, 0.10, 0.15, 0.25, 0.5]	0.1	0.1	0.15
max_depth	[**6**, 8, 10, 12, 15]	15	10	10
min_child_weight	[0.5, **1.0**, 3.0, 5.0, 7.0, 10.0]	1	3	1
subsample	[0.5, 0.6, 0.7, 0.8, 0.9, **1.0**]	1	1	1
colsample_bylevel	[0.4, 0.5, 0.6, 0.7, 0.8, 0.9, **1.0**]	0.8	0.7	0.6
colsample_bytree	[0.4, 0.5, 0.6, 0.7, 0.8, 0.9, **1.0**]	0.5	0.9	0.4
scale_pos_weight	[**1**, 3, 5, 6, 7, 9]	1	3	3
skóre		0.8015	0.8704	0.9034
čas		8:23:45	1:33:30[11]	7:47:19

V poli dostupných hodnot byly tučně zvýrazněny defaultní hodnoty implementace XGBoost. V tabulce je rovněž uvedeno skóre, kterého model s danou konfigurací dosáhl při hledání hyper-parametrů, a doba, za jak dlouho bylo nalezení hyper-parametrů provedeno. K evaluaci modelu byla použita metrika $\mathbf{F_1}$, která bude popsána v kap. 5.8.3.

Ke stanovení kvality modelu je ovšem potřeba model dále otestovat a zohlednit více metrik. Samotné skóre modelu totiž může být zavádějící.

5.8.2 Matice záměn – Confusion Matrix

Matice záměn slouží k lepšímu posouzení kvality modelu. Vysvětlení je nejlepší na příkladu. Mějme 60 záznamů spánku a 40 záznamů bdělosti. Model nám předpoví, 48 záznamů spánku správně a 12 určí jako bdělosti; dále předpoví 25 záznamů spánku správně a 15 jako bdělosti. Matice záměn pak bude následovná:

Matice záměn		Skutečnost	
– příklad		Spánek	Bdělost
Výstup modelu	Spánek	48	15
	Bdělost	12	25

Matice záměn pro binární klasifikaci využívá následující označení tříd:

- True Positive – klasifikační třída skutečně pozitivních hodnot matice záměn, model odhadl, že jev nastal, a je to pravda;
- True Negative – klasifikační třída skutečně negativních hodnot matice záměn, model odhadl, že jev nenastal, a je tomu skutečně tak;
- False Positive – klasifikační třída chybně pozitivních hodnot matice záměn, model odhadl, že jev nastal, ale ve skutečnosti nenastal;
- False Negative – klasifikační třída chybně negativních hodnot matice záměn, Model odhadl, že jev nenastal, ale mýlí se a jev nastal [2, s. 121].

Budeme li uvažovat za jev, který má nastat, spánek, pak můžeme do matice z příkladu výše přepsat třídy následujícím způsobem:

S použitím matice záměn je možné snadno definovat další metriky pro hodnocení kvality modelu. Metriky budou popsány v kap. 5.8.3.

[11] Při měření byl staticky nastaven parametr `n_estimators` a parametr `n_iter` byl snížen na 50. Díky tomu bylo hledání parametrů výrazně kratší, ovšem toto nastavení mohlo negativně ovlivnit výsledky.

Matice záměn		Skutečnost	
– třídy klasifikace		Spánek	Bdělost
Výstup modelu	Spánek	TP	FP
	Bdělost	FN	TN

5.8.3 Další metriky

Na základě matice záměn lze spočítat řadu metrik, které nám vyjadřují, v čem model vyniká a v čem naopak chybuje, a umožňují provést lepší analýzu. Následuje popis použitých metrik.

Accuracy – Přesnost

Jedná se o základní parametr, který určuje počet správných odhadů ku všem odhadům. Na nevyvážených datech může být ovšem metrika zavádějící, viz kap. 5.8.1 a zdroj [22].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Sensitivity – Senzitivita, Citlivost

Metrika **Sensitivity**, nebo též **Precision** či **True Positive Rate (TPR)**, udává poměr, v jakém je model schopný předpovědět skutečně pozitivní jev. [2, s. 122]. Aby byla metrika vypovídající, musí být hodnocena společně se svým doplňkem, metrikou Specificity V našem případě metrika značí jak dobře model rozpozná spánek.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity – Specificita

Metrika **Specificity**, či též **Recall** nebo **False Positive Rate** udává, jak je model schopný předpovědět skutečně negativní jev [2, s. 122]. Pro nás nese informaci o kvalitě předpovědí bdělého stavu subjektu.

$$\text{Specificity} = \frac{FP}{FP + TN}$$

$\mathbf{F_1}$ score

$\mathbf{F_1}$ score je metrika, která využívá obou předchozích metrik. Vychází tedy dobře pro modely, které umí správně predikovat jak skutečně pozitivní, tak skutečně negativní jev. V řadě případů může být vhodné zaměřit se na jednu z předchozích metrik

na úkor druhé, tedy maximalizovat **Sensitivity** či **Specificity** [2, p .123]. V našem případě by ovšem měl systém být schopný klasifikovat spánek i bdělost, a proto je metrika **F_1 score** velmi podstatnou veličinou.

$$\mathrm{F}_1 \text{ score} = 2 \cdot \frac{\text{Sensitivity} \cdot \text{Specificity}}{\text{Sensitivity} + \text{Specificity}}$$

Matthews Correlation Coefficient

Další zajímavou metrikou je **Matthews Correlation Coefficient (MCC)**, ačkoliv tato metrika není tolik rozšířena. Poskytuje informaci o korelaci mezi jednotlivými klasifikačními třídami. Ve statistice potom metrika odpovídá hodnotě **phi-coefficient**. Tato metrika je vhodná, pokud jsou důležité oba dva jevy [25]. V případě systému je důležitá přesnost jak spánku, tak bdění, a proto je **MCC** další užitečnou metrikou.

$$\mathrm{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN)}}$$

5.8.4 Imbalanced dataset – nevyvážený dataset

Hodnotit model pouze na základě nejzřejmnější metriky **Accuracy** může být značně zavádějící. Pokud bychom například měli dataset, který by obsahoval 90 % dat spánku a 10 % dat bdělosti, mohl by model dosahovat 90 % přesnosti odhadu jen tím, že by odhadoval vše jako spánek. Jde o problém takzvaného nevyváženého datasetu – imbalanced dataset [22].

Dataset použitý v rámci studie je taktéž nevyvážen. Obsahuje 43 034 (68,44 %) datových záznamů spánku a 19 853 (31,56 %) datových záznamů bdělosti. Dat spánku je tedy dvojnásobně více.

Jelikož bývá často optimální pracovat s vyváženým datasetem, byla v rámci **modelu C** využita metoda oversampling – nadvzorkování [22]. Byly vygenerovány syntetické hodnoty pro data v bdělém stavu. Pro vygenerování syntetických hodnot byla využita knihovna `imbalanced-learn`, konkrétně pak třída `SMOTE` využívající metody Synthetic Minority Over-sampling Technique. Trénovací dataset byl následně vybalancován, třída `SMOTE` vygenerovala syntetická data bdění taktéž pomocí stojového učení.[12]

Před samotnou operací generování syntetických dal bylo třeba odstanit hodnoty `NaN` – Not a Number, které byly při parametrizaci datasetu vytvořeny například kvůli dělení nulou, kdy kupříkladu průměr dat mohl být nulový. K odstranění hodnot

[12]Blyžší informace o knihovně `imbalanced-learn` a metodě SMOTE lze nalézt na odkaze `https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html`.

`NaN` bylo využito třídy `KNNImputer` z knihovny `scikit-learn`. Tato třída využívá k doplnění chybějících hodnot metodu k-nearest neigbours, k nejbližších sousedů, chybějící hodnota je stanovena na základě analýzy k podobných hodnot. V rámci implementace bylo k nastaveno na 4, tedy byly analyzovány 4 nejbližší hodnoty.

5.8.5 Výsledné metriky křížové validace

Metriky **Accuracy**, **Sensitivity**, **Specificity**, **F_1 score Matthews Correlation Coefficient** byly měřeny při provádění křížové validace. Je dobré zdůraznit, že křížová validace se provádí na trénovacích datech, tedy kupříkladu v případě **modelu C** na datech rozšířených o data syntetická. Výsledky jsou zobrazeny v tab. 5.2.

Tab. 5.2: Výsledky křížové validace nad jednotlivými modely

	Model A		**Model B**		**Model C**	
Metrika	Průměr	Odchylka	Průměr	Odchylka	Průměr	Odchylka
Accuracy	0,80	$\pm 0,01$	0,78	$\pm 0,00$	0,84	$\pm 0,01$
Sensitivity	0,92	$\pm 0,00$	0,96	$\pm 0,00$	0,91	$\pm 0,01$
Specificity	0,54	$\pm 0,01$	0,38	$\pm 0,01$	0,77	$\pm 0,01$
F_1	0,68	$\pm 0,00$	0,54	$\pm 0,00$	0,83	$\pm 0,01$
MCC	0,51	$\pm 0,01$	0,45	$\pm 0,01$	0,69	$\pm 0,01$

Na obr. D.1 je zobrazen graf vývoje jednotlivých metrik v průběhu procesu křížové validace pro **Model A**, na obr. D.2 pro **Model B**, a na obr. D.3 pro **Model C**. Na obrázcích D.4, D.5 a D.6 jsou dále zobrazeny metriky **Sensitivity** a **Specificity**. Obrázky naznačují, jak je model vybalancován vzhledem k připravenosti vyhodnocovat kladný děj (spánek) a záporný děj (bdění). Všechny obrázky jsou součástí příloh práce.

Nejlepších výsledků dosahuje **Model C**, u kterého je výrazně vyšší metrika **Specificity**, což se promítá do výsledků ostatních metrik. Toho je dosaženo díky vybalancování datasetu. Je ovšem třeba ověřit, jakých výsledků bude dosaženo po natrénování modelu a otestování na testovacích datech.

5.8.6 Testování natrénovaných modelů

Model byl nartrénován pomocí funkce `fit`. Této funkci byl dále předán trénovací a testovací dataset, čili byla okamžitě provedena validace. Jako validační metriky byly zvoleny metriky `error`, `logloss` a `auc`, viz kap. 5.7.5. Dataset byl rozdělen

na testovací a trénovací data v poměru 60 % ku 40 %. Matice záměn pro modely na testovacích datech jsou uvedeny v tab. 5.3, tab. 5.4 a tab. 5.5. Na základě matic záměn můžeme opět stanovit metriky, tentokrát pro testovací data. Výsledky jsou uvedeny v tab. 5.6. Jak je z tabulky patrné, všechny modely mají poměrně dobré výsledky. Výsledky jsou dokonce lepší než v případě křížové validace.

Tab. 5.3: Matice záměn na testovacích datech pro **Model A**

Matice záměn		Skutečnost	
– model A		Spánek	Bdělost
Výstup modelu	Spánek	15 854	3 609
	Bdělost	1 315	4 385

Tab. 5.4: Matice záměn na testovacích datech pro **Model B**

Matice záměn		Skutečnost	
– model B		Spánek	Bdělost
Výstup modelu	Spánek	16 588	4 897
	Bdělost	586	3 097

Tab. 5.5: Matice záměn na testovacích datech pro **Model C**

Matice záměn		Skutečnost	
– model C		Spánek	Bdělost
Výstup modelu	Spánek	16 404	4 843
	Bdělost	765	3 151

Jako nejlepší model se na testovacích datech jeví **Model A**. Zdá se tedy, že syntetická data **Modelu C** neodpovídají přesně datům testovacím.

V přílohách v kap. E jsou grafy zobrazující vývoj jednotlivých evaluačních metrik v rámci procesu učení. Dále jsou zde taktéž grafy s váhou jednotlivých parametrů, která vyjadřuje důležitost daného parametru pro rozhodování jednotlivých modelů. Zajímavým zjištěním je fakt, že všechny tři modely spoléhají především na měření teploty a méně pak na akcelerometr. **Modely A** a **B** spolehají velkou měrou na stejnou metriku, 5. percentil teploty, který v **Modelu A** zaujímá váhu asi 13 %

Tab. 5.6: Výsledky testovacích dat nad jednotlivými modely

Metrika	**Model A**	**Model B**	**Model C**
Accuracy	0,80	0,78	0,78
Sensitivity	0,92	0,97	0,95
Specificity	0,55	0,39	0,39
$\mathbf{F_1}$	0,87	0,86	0,85

a v **Modelu B** přibližně 17 %. **Model C** má váhu parametrů více rovmoměrně rozloženou, prvních osm parametrů jsou však parametry teploty a až na místě osmém je parametr akcelerometru.

Natrénované modely byly následně testovány souhrně na celém datasetu. Výsledky jsou uvedeny v tab. 5.7.

Tab. 5.7: Výsledky všech dat nad jednotlivými modely

Metrika	**Model A**	**Model B**	**Model C**
Accuracy	0,89	0,83	0,80
Sensitivity	0,97	0,97	0,98
Specificity	0,73	0,48	0,43
$\mathbf{F_1}$	0,93	0,89	0,87
MCC	0,75	0,59	0,53

Na celém datasetu má nejlepší výsledky opět **Model A**. Překvapivě jako jediný ze tří modelů dosahuje i poměrně vysoké hodnoty metriky **Specificity**, tedy by měl být schopný vyhodnocovat jak data spánku, tak data bdění dostatečně kvalitně. Z toho důvodu byl zvolen **Model A** jako finální model k implementaci.

5.9 Vizualizace výsledků

Pomocí natrénovaného modelu z předchozích kapitol lze predikovat data spánku na základě dat z aktigrafu GENEActiv. Uživatel očekává po nahrání dat na server obdržení předpovědi spánku na základě nahraných dat.

V duchu této metodiky jsou uživateli výsledné předpovědi o datech spánku zobrazeny na detailní stránce subjektu, viz obr. B.3, jako interaktivní grafy. Je možné provádět detailní analýzu díky možnosti změnit časový rozsah.

V případě dat trénovacích, tedy dat která byla součástí datasetu [15], jsou kromě dat predikce rovněž zobrazeny data polysomnografie, kdy obě varianty jsou rozlišeny barevně a popsány v legendě grafu. Zpracování nových dat trvá asi minutu na jednu noc, jednou použitá data jsou poté načtena z databáze předpovědí za přibližně 3 vteřiny.

Splněn je cíl celého zpracování dat a systému obecně, kdy uživateli jsou v srozumitelné formě prezentována data o spánku na základě dat z aktigrafu.

6 Testování implementace

Byly použity dvě metody testování, automatická a manuální. Automatickému testování byla věnována podkapitola 4.2.12.

Manuální testování celého systému mělo následující postup. S pomocí spánkového deníku, viz kap. F, byla shromažďována a dokumentována data spánku pomocí aktigrafu. Data byla následně stažena z aktigrafu pomocí aplikace GENEActivPc-Software a převedena do formátu CSV. Byla potvrzena struktura dat a funkčnost dle popisů výrobce. Připojení na webové rozhraní serveru pomocí standardního prohlížeče Google Chrome v rámci lokální sítě rovněž probíhalo bez problémů. Následně byl vytvořen účet ze skupiny výzkumných pracovníků a s jeho pomocí byla nahrána data na server a byl vytvořen příslušný subjekt.

Dalším zdrojem dat byla open source data z datasetu *Newcastle polysomnography and accelerometer data* dostupného na odkaze `https://zenodo.org/record/1160410#.XW4PBJMzb64` [14]. Tento dataset obsahuje 55 datových souborů od 28 různých subjektů. Tato data byla rovněž převedena z binární podoby do souborů CSV. Společně se soubory polysomnografie byla data nahraná na server a označena jako data trénovací. S větším množstvím dat byla otestována stabilita serveru. S takto připravenými daty teprve mohla započít fáze strojového učení.

Po natrénování modelu bylo možné vyhodnotit data, a to jak data testovací, tak data shromažďovaná a dokumentovaná spánkovým deníkem. Výsledky byly prezentovány interaktivními grafy.

Bohužel nebylo možné provést pilotní studii z důvodu epidemie nemoci Covid19. Pilotní studie ovšem bude provedena v rámci projektu NU20-04-00294 *Diagnostika onemocnění s Lewyho tělísky v prodromálním stadiu založená na analýze multimodálních dat* financovaného Ministerstvem zdravotnictví ČR. Systém je součástí tohoto projektu.

Dále byla testována stabilita serverové implementace. Byl otestován přístup na webové rozhraní s pomocí jiných prohlížečů a jiných platforem a kvalita zobrazených stránek například na mobilním telefonu.

Testování prokázalo funkčnost systému a pomohlo odhalit drobné chyby, které byly opraveny.

7 Legislativní úprava

Práce se dotýká několika legislativních otázek, které je třeba blíže popsat a zodpovědět. Tomu jsou věnovány následující podkapitoly.

7.1 Open source dataset

Pro rozšíření datové základny byl využit open source dataset *Newcastle polysomnography and accelerometer data*, který vznikl v rámci studie *Estimating sleep parameters using an accelerometer without sleep diary*. Tento dataset je dostupný pod licencí Creative Commons Attribution 4.0 International. Je tedy potřeba ověřit, zda je práce s touto licencí kompatibilní [14, 15].

Licence Creative Commons Attribution 4.0 International známá pod zkratkou CC BY 4.0 opravňuje především k „*výkonu Licencovaných práv k Licencovanému obsahu tímto způsobem a v tomto rozsahu: rozmnožování a Sdílení Licencovaného obsahu jako celku nebo jeho části,; a vytváření, rozmnožování a Sdílení Zpracovaného obsahu.*“ Aby byla licence platná, musí nabyvatel: „*zachovat následující informace, pokud je Poskytovatel s Licencovaným obsahem uvedl: identifikace tvůrce (tvůrců) Licencovaného obsahu a kohokoli dalšího, kdo má být uveden, a to jakýmkoli Poskytovatelem požadovaným způsobem (včetně pseudonymu, pokud je uveden), pokud je tento způsob rozumný; výhradu autorského práva; odkaz na tuto Veřejnou licenci; odkaz na vyloučení záruk; URI nebo hypertextový odkaz na Licencovaný obsah, pokud je to rozumně proveditelné; uvést informace o případné úpravě Licencovaného obsahu a zachovat veškeré zmínky o předchozích úpravách; a uvést, že Licencovaný obsah je licencován v souladu s touto Veřejnou licencí a uvést její text či URI nebo hypertextový odkaz na tuto Veřejnou licenci.*“ [26].

Je tedy třeba uvést korektní citaci na dataset, a to jak v práci samotné, tak i na webových stránkách. Na obou místech také musí figurovat odkaz na příslušnou licenci. Rovněž je potřeba zmínit, že data byla upravena převodem z binární podoby do agregovaných souborů CSV a následně vytěžena a zpracována v souladu s výše uvedenou licencí.

7.2 Šablona webové stránky

Pro design webové stránky byly převzaty některé styly z šablony Adequately dostupné na webových stránkách `www.templated.co`. Šablona byla upravena pro potřeby projektu. Šablona je licencována Creative Commons Attribution 3.0, známou pod zkratkou CC BY 3.0. Jedná se o starší verzi CC BY 4.0 zmíněné v podkapitole 7.1 a v klíčových právech a požadavcích je shodná s touto novější verzí [27, 28]. Je

tedy třeba uvést odkaz na původní autorské dílo a odkaz na příslušnou licenci. Tyto informace jsou uvedeny v citacích zde i na webové stránce.

7.3 Ikony

V rámci tvorby schématu 2.1 a na webových stránkách byly použity ikony ze stránky `www.flaticon.com`. Tyto ikony podléhají licenci zvěřejněné na odkaze `https://www.freepikcompany.com/legal#nav-flaticon-license`. V souladu s licencí je možné ikony použít pro nekomerční účely, ale musí být citován autor a webová stránka flaticons. Proto byly v této práci i na webové stránce uvedeny příslušné citace.

7.4 Ochrana osobních údajů

Systém slouží k získávání informací o zdravotním stavu subjektů pomocí aktigrafie. Je tedy třeba zamyslet se nad skutečností, zda pracujeme s osobními údaji a jaká jsou naše práva a povinnosti.

V rámci webového dashboardu se uvažuje o zobrazení následujících dat: kód subjektu, lokace aktigrafu, lateralita, věk, pohlaví, výška a váha. Dále pak budou uvedeny informace o zdravotním stavu vyplývající z výsledků studie.

Definice osobních údajů dle ministerstva vnitra je následující: „*Osobním údajem je každá informace o identifikované nebo identifikovatelné fyzické osobě (subjektu údajů). Identifikovatelnou fyzickou osobou je fyzická osoba, kterou lze přímo či nepřímo identifikovat, zejména odkazem na určitý identifikátor (jméno, číslo, síťový identifikátor) nebo na jeden či více zvláštních prvků fyzické, fyziologické, genetické, psychické, ekonomické, kulturní nebo společenské identity této fyzické osoby.*“ [29].

Z definice tedy vyplývá, že ačkoliv nelze jednoduše jednoznačně identifikovat subjekty studie, i tak se jedná o osobní údaje a jejich zpracování. V rámci budoucí práce tak je třeba postupovat dle platného zákona o ochraně osobních údajů.

Jelikož tato práce je akademické literární dílo, tak jest i počítačový program literárním dílem, bude relevantní legislativní úpravou druhý díl zákona 2019/110 „*Zpracování osobních údajů prováděné pro novinářské účely nebo pro účely akademického, uměleckého nebo literárního projevu*“.

Dle zákona je třeba informovat subjekty o identitě správce, což lze provést i ústně, a je třeba subjekt informovat o tom, jaké údaje budou zpracovávány a za jakým účelem. Subjekt musí mít dále možnost přistoupit ke svým údajům, které jsou zpracovávány. Pokud by subjekt žádal o výmaz dat, má na to právo pouze tehdy, nepotřebuje-li správce již osobní údaje pro účely zpracování [30, § 17–23]. Podrobné informace

lze dohledat v citovaném zákoně a v nařízení Evropského parlamentu a Rady (EU) 2016/679 [31].

Pokud by bylo shledáno, že se v rámci práce jedná o případné zpracování za účelem vědecké činnosti, bylo by třeba splnit navíc požadavky dle § 16 zákona 2019/110. Požadavky jsou například:

- vedení záznamů o všech operacích s osobními údaji a jejich uchování po dobu dvou let,
- omezení přístupu k osobním údajům v rámci správce nebo zpracovatele,
- pseudonymizace osobních údajů,
- šifrování osobních údajů,
- opatření k zajištění trvalé důvěrnosti, integrity, dostupnosti a odolnosti systému a služeb zpracování,
- proces pravidelného testování a podobně [30, § 16].

I v rámci těchto přísnějších požadavků by mělo být implementované řešení vyhovující. Operace jsou zaznamenány pomocí logů, které lze uchovávat po libovolně dlouhou dobu, přístup je omezen pouze na uživatele s administrátorským právem, uživatelé mají možnost prohlédnout si své osobní údaje, které jsou pseudonymizovány s pomocí identifikačního kódu. Rovněž je zajištěno šifrování databáze a použity mechanizmy pro zajištění důvěrnosti, integrity, dostupnosti a odolnosti.

Na závěr je třeba zdůraznit fakt, že v rámci diplomové práce nebyly zpracovávány žádné osobní údaje. K testování byly použity pouze pseudonymizovaná data z open source datasetu, viz kapitola 7.1. Rovněž nedošlo k zveřejnění žádných dat, neboť server je pro účely diplomové práce dostupný pouze na lokální síti.

8 Závěr

Cílem práce bylo navrhnout a implementovat systém zabezpečeného přenosu a zpracování dat z aktigrafu. Navržený a realizovaný systém byl v rámci diplomové práce podrobně představen a popsán a následně byla ověřena jeho funkčnost na datasetu velikosti obvyklé studie. Podrobně byla rovněž analyzována a představena bezpečnost systému. Následně byly implementovány metody strojového učení a povedlo se natrénovat model, který dokáže rozpoznat, zda subjekt spí či bdí na základě dat z aktigrafu. Systém byl shledán plně funkčním pro potřeby diplomové práce.

Rovněž byl před samotnou realizací prozkoumán stav techniky, byly představeny možné alternativy a bylo zvoleno výsledné řešení. Naplněn byl taktéž cíl prověřit příslušnou legislativní úpravu a zajistit nezávadnost systému v rámci zákonných požadavků, pro použití v reálné studii. Řešení bylo patřičně přizpůsobeno platné legislativě. Rovněž byly splněny právní požadavky, které byly uloženy tvůrci externích doplňků a materiálů v rámci jejich zákonných práv.

Systém celkově prokázal svoji funkčnost a kompatibilitu s legislativní úpravou, taktéž prokázal vhodnost využití v rámci projektu NU20-04-00294 *Diagnostika onemocnění s Lewyho tělísky v prodromálním stadiu založená na analýze multimodálních dat*. Při implementaci a návrhu systému byla brána v potaz co největší dostupnost cílové skupině uživatelů a jednoduchá použitelnost. Výsledkem práce je funkční systém navržený plně dle trendu Health 4.0.

Systém vytvořený v rámci této diplomové práce bude využíván ve Středoevropském technologickém institutu a ve Fakultní nemocnici u sv. Anny v Brně k výzkumu poruch spánku u osob, u kterých je vysoké riziko onemocnění s Lewyho tělísky. Systém tak napomůže vytvořit metodologii, která umožní diagnostikovat tato onemocnění v raném stádiu.

Literatura

[1] MORGENTHALER, T., et al. *Practical parameters for the Use of Actigraphy in the Assessmnet of Sleep and Sleep Disorders:An Update for 2007* [online]. Sleep, vol. 30, Issue 4. 2007 [cit. 21. 9. 2018]. Dostupné z URL: <https://academic.oup.com/sleep/article/30/4/519/2708218>.

[2] GÉRON, A. *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems.* Tokyo: O'Reilly, [2017]. ISBN 978-1-4919-6229-9.

[3] QUANTE, M., et al. *Actigraphy-Based Sleep Estimation in Adolescents and Adults: A Comparison with Polysomnography Using Two Scoring Algorithms.* [online]. Nature and Science of Sleep 10, 2018 [cit. 21. 9. 2018]. Dostupné z URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5779275/>.

[4] Activinsights: *Publications* [online]. Cambridgeshire: 2017, poslední aktualizace květen 2017 [cit. 24. 2. 2019]. Dostupné z URL: <https://49wvycy00mv416l561vrj345-wpengine.netdna-ssl.com/wp-content/uploads/2017/07/Activinsights-Publications-May-2017.pdf>.

[5] HEESCH, K. C., et al. *Validity of objective methods for measuring sedentary behaviour in older adults: a systematic review* [online]. International Journal of Behavioral Nutrition and Physical Activity, 2018 15:199 [cit. 24. 2. 2019]. Dostupné z URL: <https://ijbnpa.biomedcentral.com/track/pdf/10.1186/s12966-018-0749-2>.

[6] BURGET, P. *Teoretická informatika*, Brno: Vysoké učení technické v Brně, 2013. ISBN 978-80-214-4897-1.

[7] Activinsights: *GENEActiv Instruction Manual v 1.2* [online]. Cambridgeshire: 2014, [cit. 3. 3. 2019]. Dostupné z URL: <https://49wvycy00mv416l561vrj345-wpengine.netdna-ssl.com/wp-content/uploads/2014/03/geneactiv_instruction_manual_v1.2.pdf>.

[8] WHITE, T. *Thomite/pampro v0.4.0 (Version v0.4.0).* Zenodo:2018, poslední aktualizace 2. 3. 2018 [cit. 3. 3. 2019]. Dostupné z URL: <http://doi.org/10.5281/zenodo.1187043>.

[9] SMITH, M. L.; ERWIN, J.; DIAFERIO, S. *Role & responsibility charting (RACI)*[online]. Project Management Forum. 2005 [cit. 18. 10. 2019]. Dostupné

z URL:
<https://pmicie.starchapter.com/images/downloads/raci_r_web3_1.pdf>.

[10] OWASP: *SCG WF Django* [online]. Secure configuration guide. Poslední aktualizace 26. 2. 2015 [cit. 3. 11. 2019]. Dostupné z URL:
<https://www.owasp.org/index.php/SCG_WF_Django>.

[11] Django: *Deploying Django*[online]. [cit. 16. 12. 2019]. Dostupné z URL:
<https://docs.djangoproject.com/en/3.0/howto/deployment/>.

[12] TRIPLETT, J.; HENSCHEL, L. W. *10 tips for making the Django Admin more secure*[online]. Opensource.com, 2018, poslední aktualizace 31. 1. 2018 [cit. 11. 12. 2019]. Dostupné z URL:
<https://opensource.com/article/18/1/10-tips-making-django-admin-more-secure>.

[13] Django: *Security in Django*[online]. [cit. 14. 12. 2019]. Dostupné z URL:
<https://docs.djangoproject.com/en/3.0/topics/security/>.

[14] VAN HEES, V. T., et al. *Newcastle polysomnography and accelerometer data*[online]. Zenodo, poslední aktualizace 25. 1. 2018 [cit. 1. 11. 2019]. Dostupné z URL:
<https://doi.org/10.5281/zenodo.1160410>.

[15] VAN HEES, V. T., et al. *Estimating sleep parameters using an accelerometer without sleep diary*[online]. Scientific reports, 2018, 8.1: 12975. [cit. 1. 11. 2019]. Dostupné z URL:
<https://doi.org/10.1038/s41598-018-31266-z>.

[16] MORÁŇ, M. *Poruchy spánku* Solen, 2001, 1: 2NREM. Dostupné z URL:
<http://www.solen.cz/pdfs/int/2001/03/02.pdf>.

[17] KRÁLÍK, R., et al. *Klasifikace spánkových fází pomocí PSG dat.* Elektrorevue, 2016, 18.6. ISSN 1213-1539. Dostupné z URL:
<http://www.elektrorevue.cz/cz/clanky/zpracovani-signalu/10/klasifikace-spankovych-fazi-pomoci-psg-dat--classification-of-sleep-stages-using-psg-data-/>.

[18] FARRAHI, V., et al. *Calibration and validation of accelerometer-based activity monitors: A systematic review of machine-learning approaches.* Gait & posture, 2019, 68: 285-299. Dostupné z URL:

<https://www.sciencedirect.com/science/article/pii/S0966636218313092#bib0230>.

[19] CHEN, T., GUESTRIN, C. *Xgboost: A scalable tree boosting system.* Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, s. 785-794. Dostupné z URL: <https://dl.acm.org/doi/abs/10.1145/2939672.2939785>.

[20] JŮZLOVÁ, M. *Model Performance Approximation in Hyper-parameter Optimization.* Master's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2018. Dostupné z URL: <http://hdl.handle.net/10467/76421>.

[21] XGBoost: *XGBoost Parameters.* [online] [cit. 22. 5. 2020]. Dostupné z URL: <https://xgboost.readthedocs.io/en/latest/parameter.html>.

[22] BOYLE, T. *Dealing with Imabalanced Data* [online]. towards data science: 2019, poslední aktualizace 3. 2. 2019 [cit. 23. 5. 2020]. Dostupné z URL: <https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18>.

[23] TUAN, H. A., *Binární klasifikace pomocí rozhodovacích stromů* [online]. Diplomová práce. Vysoká škola ekonomická v Praze. Praha, 2018 [cit. 2020-05-23]. Dostupné z: <https://theses.cz/id/3j2j6k/>.

[24] scikit-learn: *Cross-validation: evaluating estimator performance* [online]. [cit. 24. 5. 2020]. Dostupné z URL: <https://scikit-learn.org/stable/modules/cross_validation.html>.

[25] SHMUELI, B. *Matthews Correlation Coefficient is The Best Classification Metric You've Never Heard Of* [online]. towards data science: 2019, poslední aktualizace 22. 11. 2019 [cit. 27. 5. 2020]. Dostupné z URL: <https://towardsdatascience.com/the-best-classification-metric-youve-never-heard-of-the-matthews-correlation-coefficient-3bf50a2f3e9a>.

[26] Creative Commons: *Attribution 4.0 International Public License*[online]. [cit. 12. 12. 2019]. Dostupné z URL: <https://creativecommons.org/licenses/by/4.0/legalcode>.

[27] Templated: *Adequately*[online]. [cit. 13. 12. 2019]. Dostupné z URL: <https://templated.co/adequately>.

[28] Creative Commons: *Attribution 3.0 Unported*[online]. [cit. 13. 12. 2019]. Dostupné z URL:
<https://creativecommons.org/licenses/by/3.0/>.

[29] Ministerstvo vnitra České republiky: *Základní pojmy v GDPR*[online]. 2019, [cit. 16. 12. 2019]. Dostupné z URL:
<https://www.mvcr.cz/gdpr/clanek/zakladni-pojmy-v-gdpr.aspx>.

[30] Zákon č. 110/2019 Sb.: *Zákon o zpracování osobních údajů*[online]. [cit. 16. 12. 2019]. Dostupné z URL:
<https://www.zakonyprolidi.cz/cs/2019-110>.

[31] *Nařízení Evropského parlamentu a rady (EU) 2016/679 ze dne 27. dubna 2016 o ochraně fyzických osob v souvislosti se zpracováním osobních údajů a o volném pohybu těchto údajů a o zrušení směrnice 95/46/ES (obecné nařízení o ochraně osobních údajů)*[online]. [cit. 16. 12. 2019]. Dostupné z URL:
<https://www.uoou.cz/assets/File.ashx?id_org=200144&id_dokumenty=20112>.

Seznam symbolů, veličin a zkratek

ANN	Artificial Neural Network – umělá neuronová síť
API	Application Programming Interface – set metod, protokolů a definic pro tvorbu odvozených aplikací
CC BY 3.0	Creative Commons Attribution 3.0 – licence pro sdílení díla s podmínkou citace
CC BY 4.0	Creative Commons Attribution 4.0 – licence pro sdílení díla s podmínkou citace
CSRF	Cross Site Request Forgery – zneužití uživatelských oprávnění autentizovaného uživatele k vykonání nechtěných akcí bez vědomí uživatele
CSS	Cross Site Scripting – vložení skriptu z cizího zdroje, který je následně vykonán prohlížečem na straně uživatele
CSV	Comma Separated Values – formát dat oddělených separátorem, obvykle čárkou nebo středníkem
DT	Decision Tree – rozhodovací strom
FN	False Negative – klasifikační třída chybně negativních hodnot matice záměn
FP	False Positive – klasifikační třída chybně pozitivních hodnot matice záměn
LR	Linear regression – lineární regrese
MCC	Matthews Corrrelation Coefficient – skalární metrika pro měření kvality binární klasifikace
OS	Operační Systém – základní software zařízení
OWASP	Open Web Application Security Project – nadace zaměřená na bezpečnost webových aplikací
RF	Random Forest – náhodný les
SVM	Support Vector Machines – metoda podpůrných vektorů
TN	True Negative – klasifikační třída skutečně negativních hodnot matice záměn
TP	True Positive – klasifikační třída skutečně pozitivních hodnot matice záměn

Seznam příloh

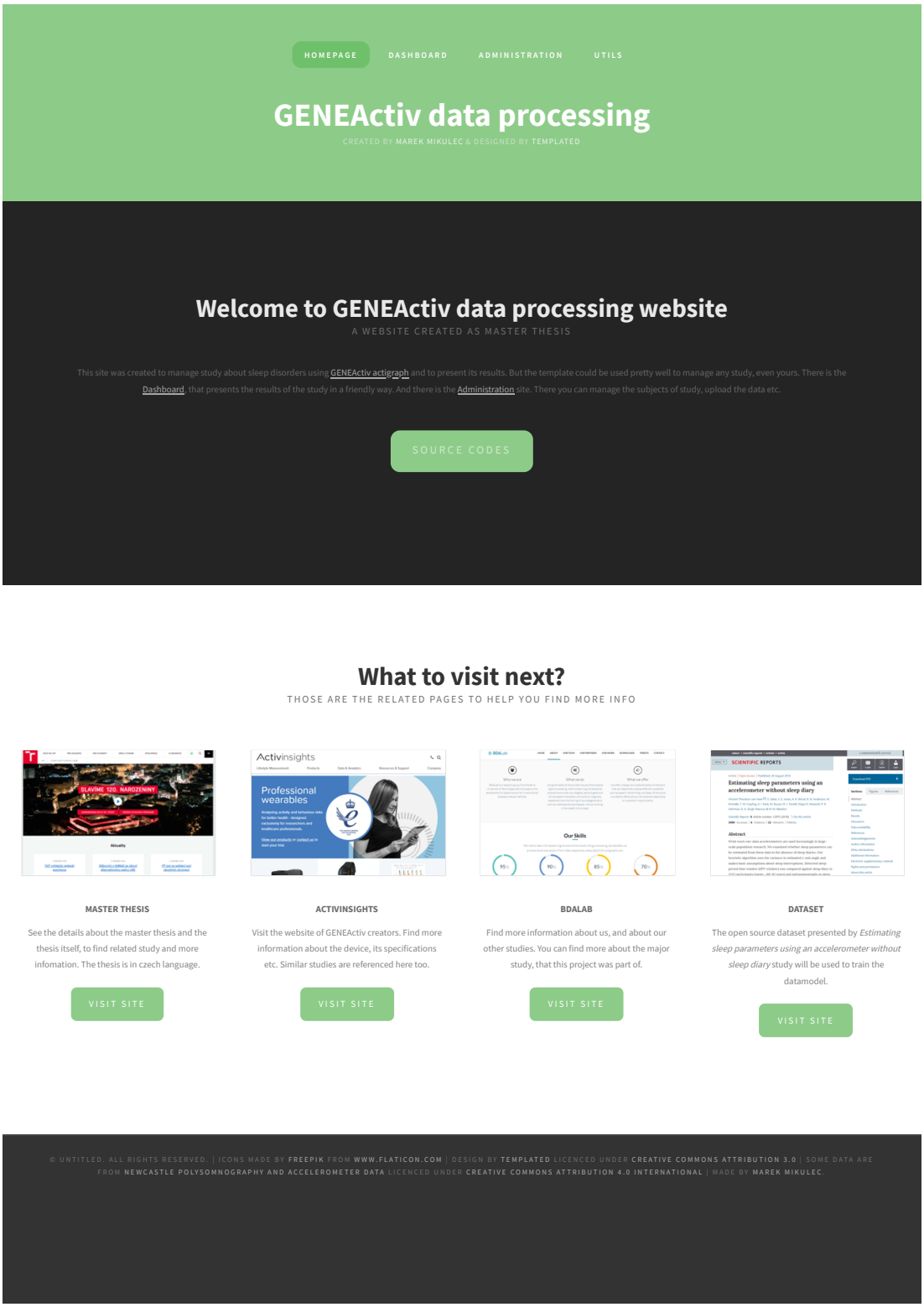
A Hardwarové parametry serveru

Jako server byl použit herní notebook Acer Aspire V15 Nitro Black edition. Níže následuje víčet hardwarových parametrů tohoto zařízení.

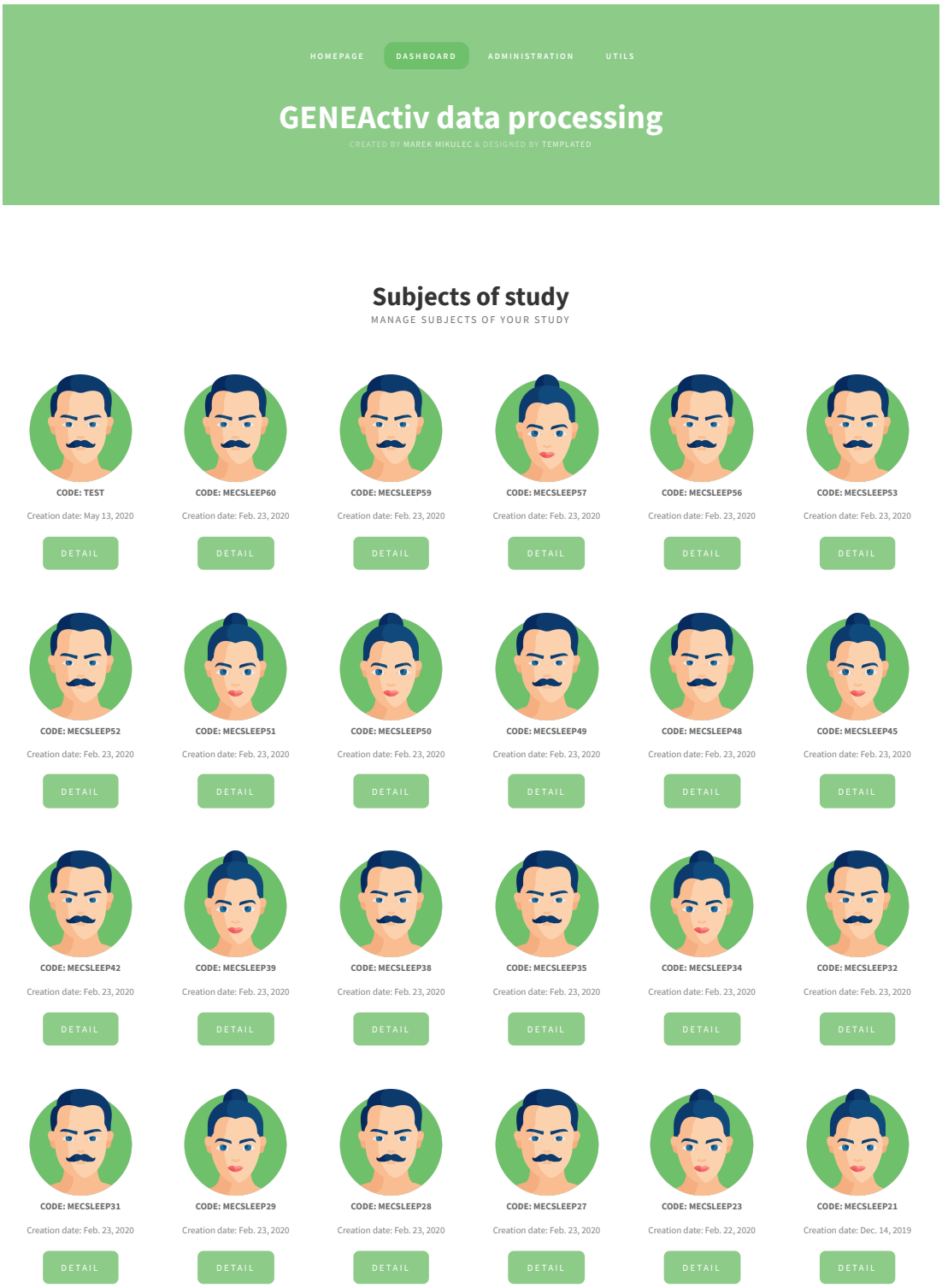
Tab. A.1: Hardwarové parametry serveru

Komponenta	**Název**	**Hodnota**
Procesor	Intel Core i7 4710HQ	$4 \times 3,5$ GHz
Grafická karta	NVIDIA GeForce GTX 860M	2 GB
RAM	DDR3	16 GB
Disk 1	SSD	128 GB
Disk 2	HDD	1000 GB
Systém	Windows 10	64 b
VM systém	Fedora 31	64 b
Rok výroby		2014

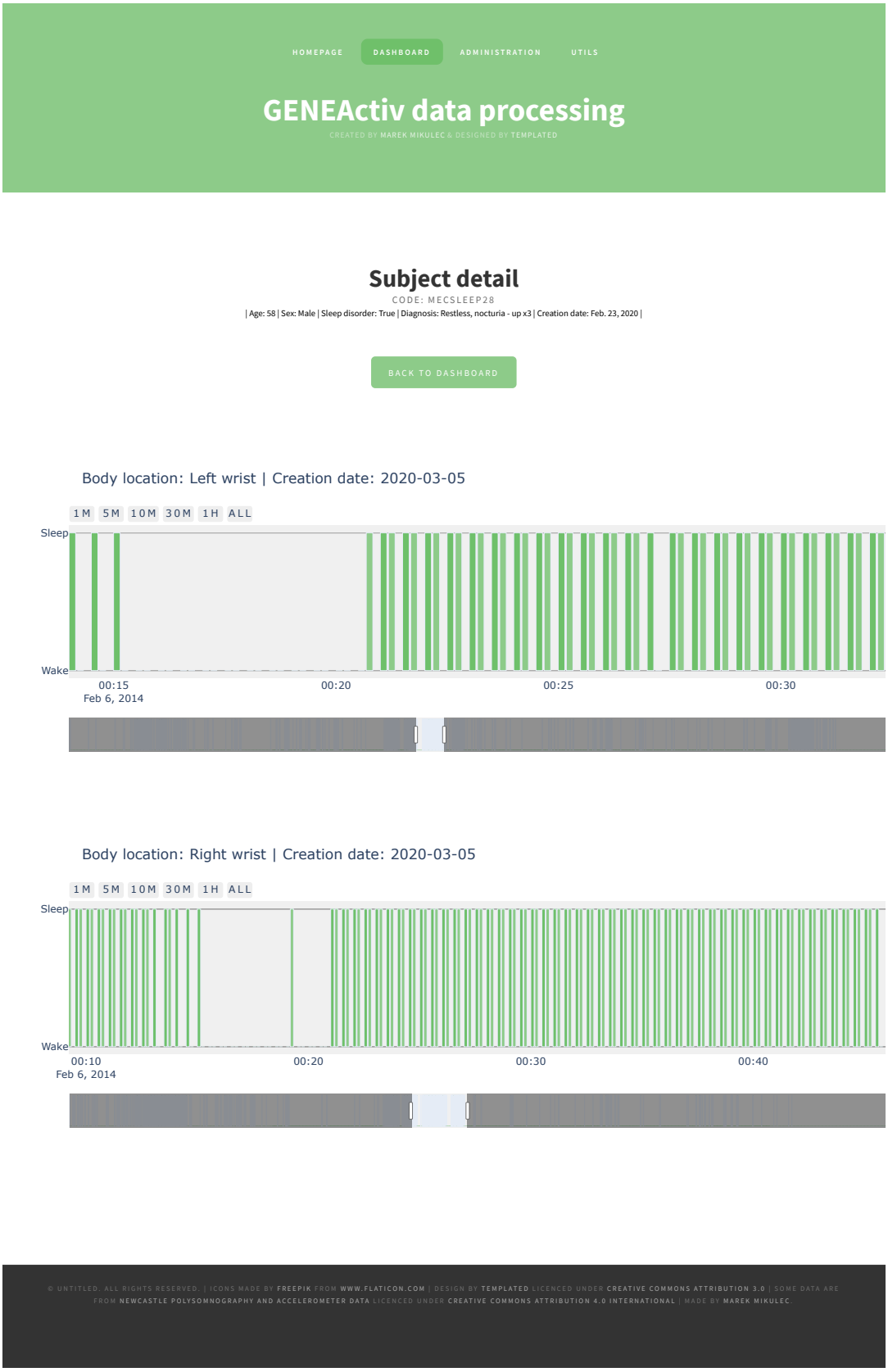
B Vzhled webového rozhraní serveru



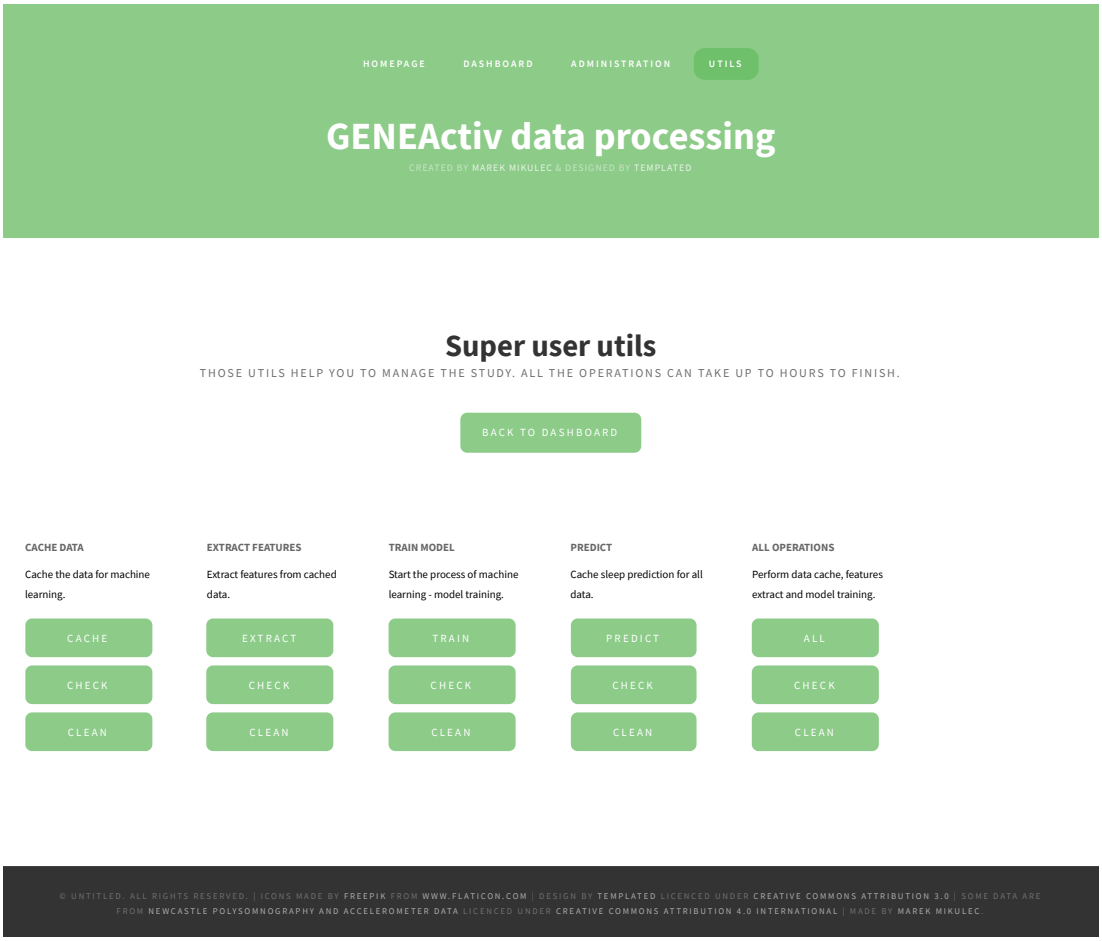
Obr. B.1: Domovská stránka, exportováno do formátu pdf



Obr. B.2: Stránka dashboard: subjekty, exportováno do formátu pdf



Obr. B.3: Stránka dashboard: detail, exportováno do formátu pdf, oříznuto



Obr. B.4: Stránka utilit, exportováno do formátu pdf

C Statistické parametry dat

C.1 Význam parametrů

Tab. C.1: Extrahované parametry dat

	Parametr	**Název ve zdrojovém kódu**
1	Maximum	MAX
2	Minimum	MIN
3	Relativní pozice maxima	RELATIVE POSITION OF MAX
4	Relativní pozice minima	RELATIVE POSITION OF MIN
5	Rozsah	RANGE
6	Rozsah relativní vůči maximu	RELATIVE RANGE
7	Relativní rozsah	RELATIVE VARIATION RANGE
8	Mezikvartilový rozsah	INTERQUARTILE RANGE
9	Relativní mezikvartilový rozsah	RELATIVE INTERQUARTILE RANGE
10	Mezidecicolový rozsah	INTERDECILE RANGE
11	Relativní mezidecicolový rozsah	RELATIVE INTERDECILE RANGE
12	Rozsah mezi 1. a 99. percentilem	INTERPERCENTILE RANGE
13	Relativní rozsah mezi 1. a 99. percentilem	RELATIVE INTERPERCENTILE RANGE
14	Studentizovaný rozsah	STUDENTIZED RANGE
15	Průměr	MEAN
16	Průměr bez 10 % odlehlých hodnot	MEAN EXCLUDING OUTLIERS (10)
17	Průměr bez 20 % odlehlých hodnot	MEAN EXCLUDING OUTLIERS (20)
18	Průměr bez 30 % odlehlých hodnot	MEAN EXCLUDING OUTLIERS (30)
19	Průměr bez 40 % odlehlých hodnot	MEAN EXCLUDING OUTLIERS (40)
20	Průměr bez 50 % odlehlých hodnot	MEAN EXCLUDING OUTLIERS (50)
21	Medián	MEDIAN
22	Modus	MODE
23	Rozptyl	VARIANCE
24	Směrodatná odchylka	STANDARD DEVIATION
25	Střední absolutní odchylka	MEDIAN ABSOLUTE DEVIATION
26	Relativní směrodatná odchylka	RELATIVE STANDARD DEVIATION
27	Index disperse	INDEX OF DISPERSION
28	Špičatost	KURTOSIS

29	Šikmost	SKEWNESS
30	Pearsonův 1. koeficient šikmosti	PEARSONS 1st SKEWNESS COEFFICIENT
31	Pearsonův 2. koeficient šikmosti	PEARSONS 2nd SKEWNESS COEFFICIENT
32	1. percentil	1st PERCENTILE
33	5. percentil	5th PERCENTILE
34	10. percentil	10th PERCENTILE
35	20. percentil	20th PERCENTILE
36	1. kvartil	1st QUARTILE
37	30. percentil	30th PERCENTILE
38	40. percentil	40th PERCENTILE
39	60. percentil	60th PERCENTILE
40	70. percentil	70th PERCENTILE
41	3. kvartil	3th QUARTILE
42	80. percentil	80th PERCENTILE
43	90. percentil	90th PERCENTILE
44	95. percentil	95th PERCENTILE
45	99. percentil	99th PERCENTILE

V rámci diplomové práce byly tyto charakteristiky stanoveny pro

- vektor dat akcelerometru
- vektor dat teploty.

Názvy veličin jsou tedy rozlišeny v uživatelském rozhraní jako

- ACCELEROMETER | VELIČINA, např. ACCELEROMETER | MEDIAN
- TEMPERATURE | VELIČINA, např. TEMPERATURE | MEAN.

C.2 Informace o kompletním datasetu

```
DatetimeIndex: 62907 entries, 2013-06-10 21:22:40 to 2015-05-21 06:20:49
Data columns (total 91 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   ACCELEROMETER | MAX                            62907 non-null  float64
 1   ACCELEROMETER | MIN                            62907 non-null  float64
 2   ACCELEROMETER | RELATIVE POSITION OF MAX       62907 non-null  float64
 3   ACCELEROMETER | RELATIVE POSITION OF MIN       62907 non-null  float64
 4   ACCELEROMETER | RANGE                          62907 non-null  float64
 5   ACCELEROMETER | RELATIVE RANGE                 62898 non-null  float64
 6   ACCELEROMETER | RELATIVE VARIATION RANGE       62907 non-null  float64
 7   ACCELEROMETER | INTERQUARTILE RANGE            62907 non-null  float64
 8   ACCELEROMETER | RELATIVE INTERQUARTILE RANGE   62898 non-null  float64
```

```
14  9   ACCELEROMETER | INTERDECILE RANGE                        62907 non-null  float64
15  10  ACCELEROMETER | RELATIVE INTERDECILE RANGE               62898 non-null  float64
16  11  ACCELEROMETER | INTERPERCENTILE RANGE                    62907 non-null  float64
17  12  ACCELEROMETER | RELATIVE INTERPERCENTILE RANGE           62898 non-null  float64
18  13  ACCELEROMETER | STUDENTIZED RANGE                        62907 non-null  float64
19  14  ACCELEROMETER | MEAN                                     62907 non-null  float64
20  15  ACCELEROMETER | MEAN EXCLUDING OUTLIERS (10)             62907 non-null  float64
21  16  ACCELEROMETER | MEAN EXCLUDING OUTLIERS (20)             62907 non-null  float64
22  17  ACCELEROMETER | MEAN EXCLUDING OUTLIERS (30)             62907 non-null  float64
23  18  ACCELEROMETER | MEAN EXCLUDING OUTLIERS (40)             62907 non-null  float64
24  19  ACCELEROMETER | MEAN EXCLUDING OUTLIERS (50)             53882 non-null  float64
25  20  ACCELEROMETER | MEDIAN                                   62907 non-null  float64
26  21  ACCELEROMETER | MODE                                     62907 non-null  float64
27  22  ACCELEROMETER | VARIANCE                                 62907 non-null  float64
28  23  ACCELEROMETER | STANDARD DEVIATION                      62907 non-null  float64
29  24  ACCELEROMETER | MEDIAN ABSOLUTE DEVIATION               62907 non-null  float64
30  25  ACCELEROMETER | RELATIVE STANDARD DEVIATION             62907 non-null  float64
31  26  ACCELEROMETER | INDEX OF DISPERSION                     62907 non-null  float64
32  27  ACCELEROMETER | KURTOSIS                                 62907 non-null  float64
33  28  ACCELEROMETER | SKEWNESS                                 62907 non-null  float64
34  29  ACCELEROMETER | PEARSONS 1st SKEWNESS COEFFICIENT       62907 non-null  float64
35  30  ACCELEROMETER | PEARSONS 2nd SKEWNESS COEFFICIENT       62907 non-null  float64
36  31  ACCELEROMETER | 1st PERCENTILE                          62907 non-null  float64
37  32  ACCELEROMETER | 5th PERCENTILE                          62907 non-null  float64
38  33  ACCELEROMETER | 10th PERCENTILE                         62907 non-null  float64
39  34  ACCELEROMETER | 20th PERCENTILE                         62907 non-null  float64
40  35  ACCELEROMETER | 1st QUARTILE                           62907 non-null  float64
41  36  ACCELEROMETER | 30th PERCENTILE                         62907 non-null  float64
42  37  ACCELEROMETER | 40th PERCENTILE                         62907 non-null  float64
43  38  ACCELEROMETER | 60th PERCENTILE                         62907 non-null  float64
44  39  ACCELEROMETER | 70th PERCENTILE                         62907 non-null  float64
45  40  ACCELEROMETER | 3th QUARTILE                           62907 non-null  float64
46  41  ACCELEROMETER | 80th PERCENTILE                         62907 non-null  float64
47  42  ACCELEROMETER | 90th PERCENTILE                         62907 non-null  float64
48  43  ACCELEROMETER | 95th PERCENTILE                         62907 non-null  float64
49  44  ACCELEROMETER | 99th PERCENTILE                         62907 non-null  float64
50  45  TEMPERATURE | MAX                                       62907 non-null  float64
51  46  TEMPERATURE | MIN                                       62907 non-null  float64
52  47  TEMPERATURE | RELATIVE POSITION OF MAX                  62907 non-null  float64
53  48  TEMPERATURE | RELATIVE POSITION OF MIN                  62907 non-null  float64
54  49  TEMPERATURE | RANGE                                     62907 non-null  float64
55  50  TEMPERATURE | RELATIVE RANGE                           62907 non-null  float64
56  51  TEMPERATURE | RELATIVE VARIATION RANGE                  62907 non-null  float64
57  52  TEMPERATURE | INTERQUARTILE RANGE                       62907 non-null  float64
58  53  TEMPERATURE | RELATIVE INTERQUARTILE RANGE              62907 non-null  float64
59  54  TEMPERATURE | INTERDECILE RANGE                        62907 non-null  float64
60  55  TEMPERATURE | RELATIVE INTERDECILE RANGE               62907 non-null  float64
61  56  TEMPERATURE | INTERPERCENTILE RANGE                    62907 non-null  float64
62  57  TEMPERATURE | RELATIVE INTERPERCENTILE RANGE           62907 non-null  float64
63  58  TEMPERATURE | STUDENTIZED RANGE                        23305 non-null  float64
64  59  TEMPERATURE | MEAN                                     62907 non-null  float64
65  60  TEMPERATURE | MEAN EXCLUDING OUTLIERS (10)             62907 non-null  float64
66  61  TEMPERATURE | MEAN EXCLUDING OUTLIERS (20)             62907 non-null  float64
67  62  TEMPERATURE | MEAN EXCLUDING OUTLIERS (30)             62907 non-null  float64
68  63  TEMPERATURE | MEAN EXCLUDING OUTLIERS (40)             62907 non-null  float64
69  64  TEMPERATURE | MEAN EXCLUDING OUTLIERS (50)             53882 non-null  float64
70  65  TEMPERATURE | MEDIAN                                   62907 non-null  float64
71  66  TEMPERATURE | MODE                                     62907 non-null  float64
72  67  TEMPERATURE | VARIANCE                                 62907 non-null  float64
```

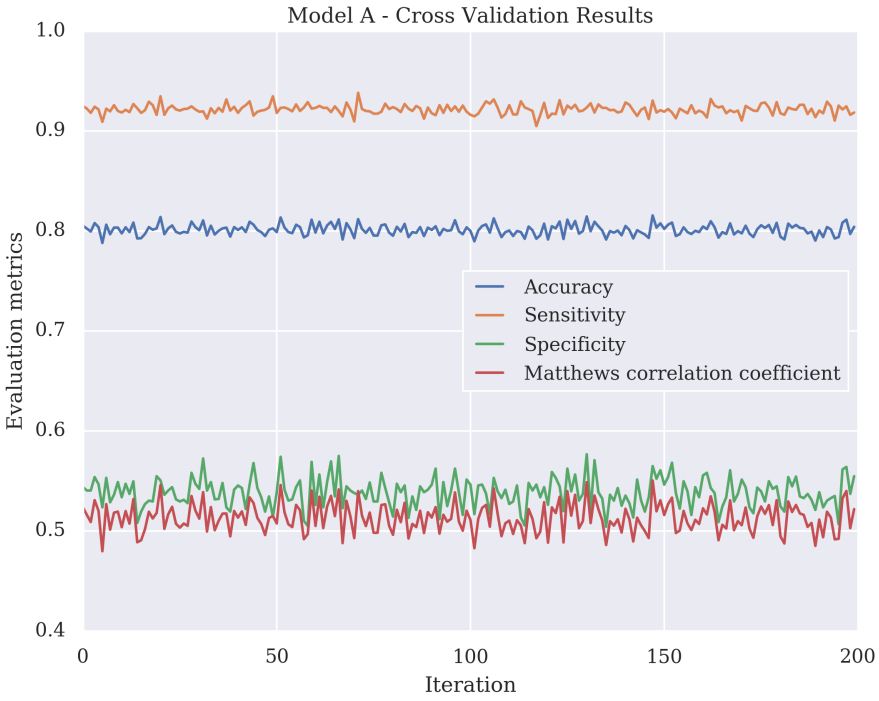
```
 68  TEMPERATURE | STANDARD DEVIATION                          62907 non-null  float64
 69  TEMPERATURE | MEDIAN ABSOLUTE DEVIATION                  62907 non-null  float64
 70  TEMPERATURE | RELATIVE STANDARD DEVIATION                62907 non-null  float64
 71  TEMPERATURE | INDEX OF DISPERSION                        62907 non-null  float64
 72  TEMPERATURE | KURTOSIS                                   62907 non-null  float64
 73  TEMPERATURE | SKEWNESS                                   62907 non-null  float64
 74  TEMPERATURE | PEARSONS 1st SKEWNESS COEFFICIENT          23305 non-null  float64
 75  TEMPERATURE | PEARSONS 2nd SKEWNESS COEFFICIENT          23305 non-null  float64
 76  TEMPERATURE | 1st PERCENTILE                             62907 non-null  float64
 77  TEMPERATURE | 5th PERCENTILE                             62907 non-null  float64
 78  TEMPERATURE | 10th PERCENTILE                            62907 non-null  float64
 79  TEMPERATURE | 20th PERCENTILE                            62907 non-null  float64
 80  TEMPERATURE | 1st QUARTILE                               62907 non-null  float64
 81  TEMPERATURE | 30th PERCENTILE                            62907 non-null  float64
 82  TEMPERATURE | 40th PERCENTILE                            62907 non-null  float64
 83  TEMPERATURE | 60th PERCENTILE                            62907 non-null  float64
 84  TEMPERATURE | 70th PERCENTILE                            62907 non-null  float64
 85  TEMPERATURE | 3th QUARTILE                               62907 non-null  float64
 86  TEMPERATURE | 80th PERCENTILE                            62907 non-null  float64
 87  TEMPERATURE | 90th PERCENTILE                            62907 non-null  float64
 88  TEMPERATURE | 95th PERCENTILE                            62907 non-null  float64
 89  TEMPERATURE | 99th PERCENTILE                            62907 non-null  float64
 90  SLEEP                                                    62907 non-null  int64
dtypes: float64(90), int64(1)
memory usage: 44.2 MB
```

Výpis C.1: Specifikace datasetu

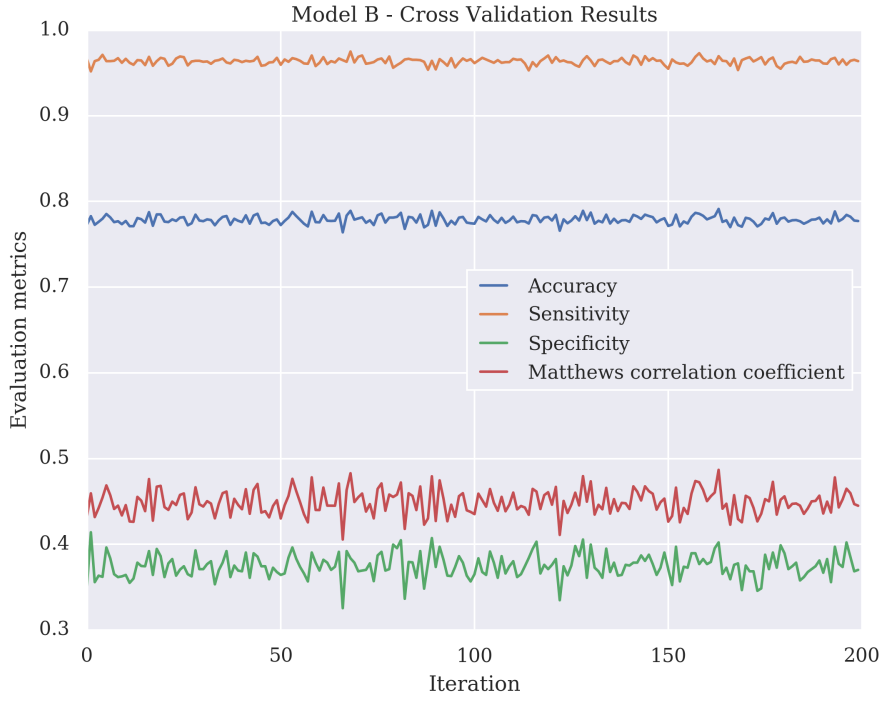
D Výsledky křížové validace

Metriky byly měřeny v průběhu křížové validace, na ose Y je hodnota jednotlivých koeficientů, na ose X se pohybujeme jednotlivými iteracemi křížové validace, tedy se pohybujeme jednotlivými dávkami dat.

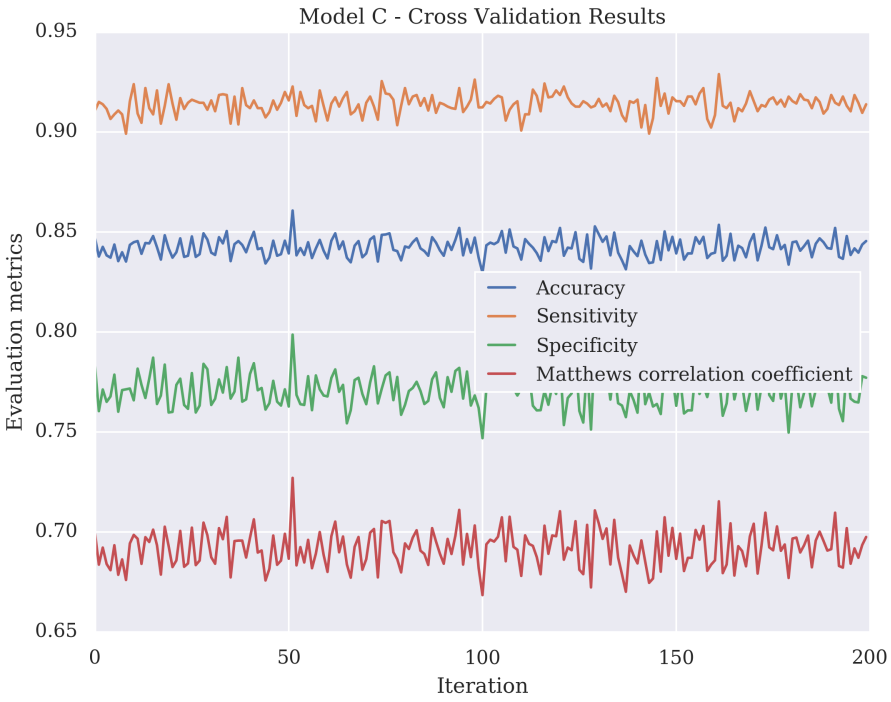
Další grafy poté zobrazují rozložení poměru mezi metrikou **Sensitivity** a metrikou **Specificity** v průběhu křížové validace.



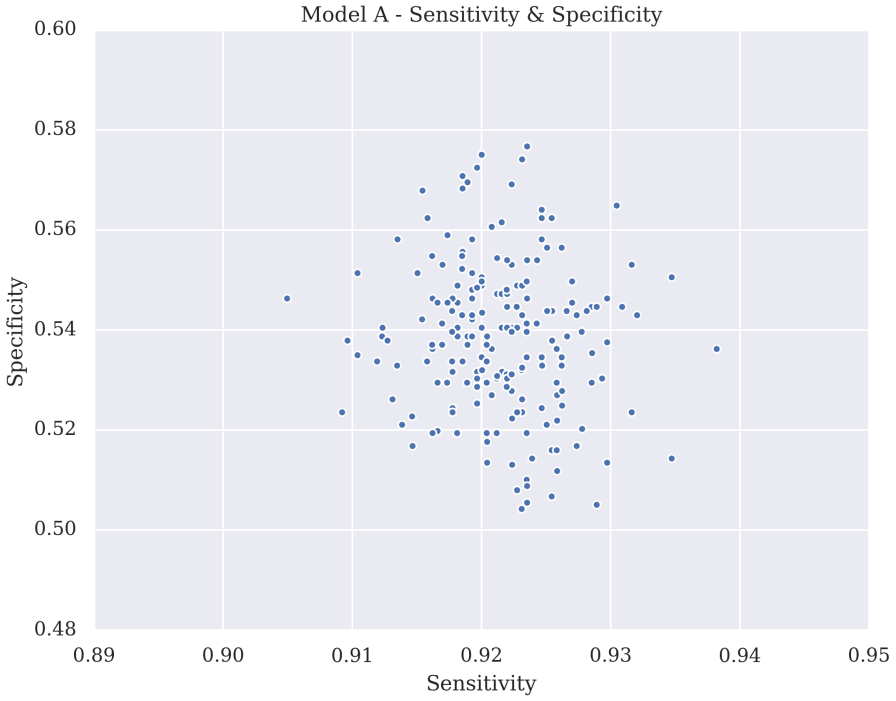
Obr. D.1: Výsledky křížové validace pro Model A



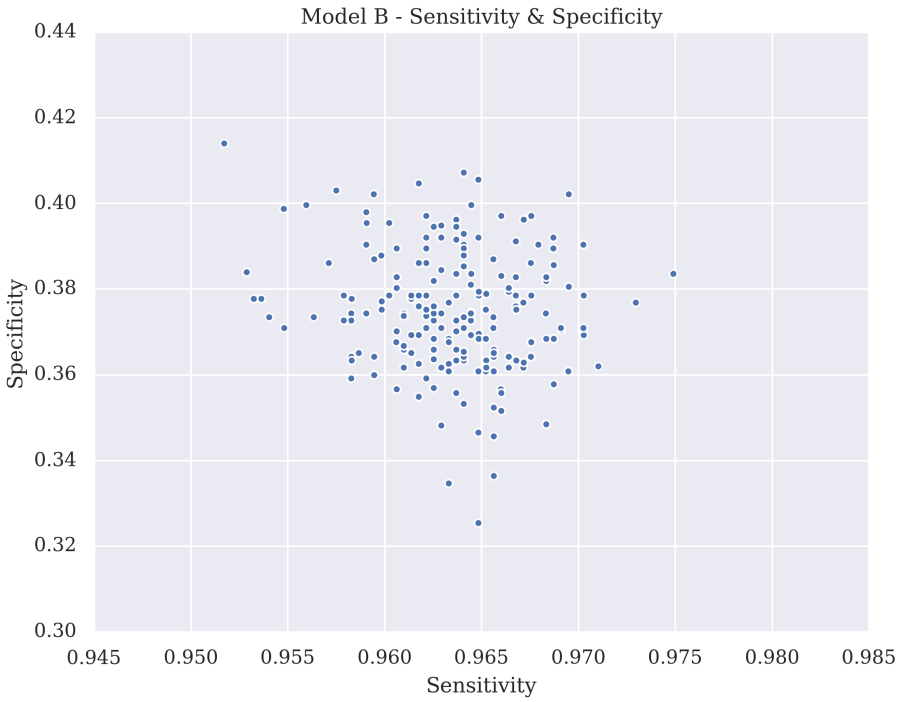
Obr. D.2: Výsledky křížové validace model pro Model B



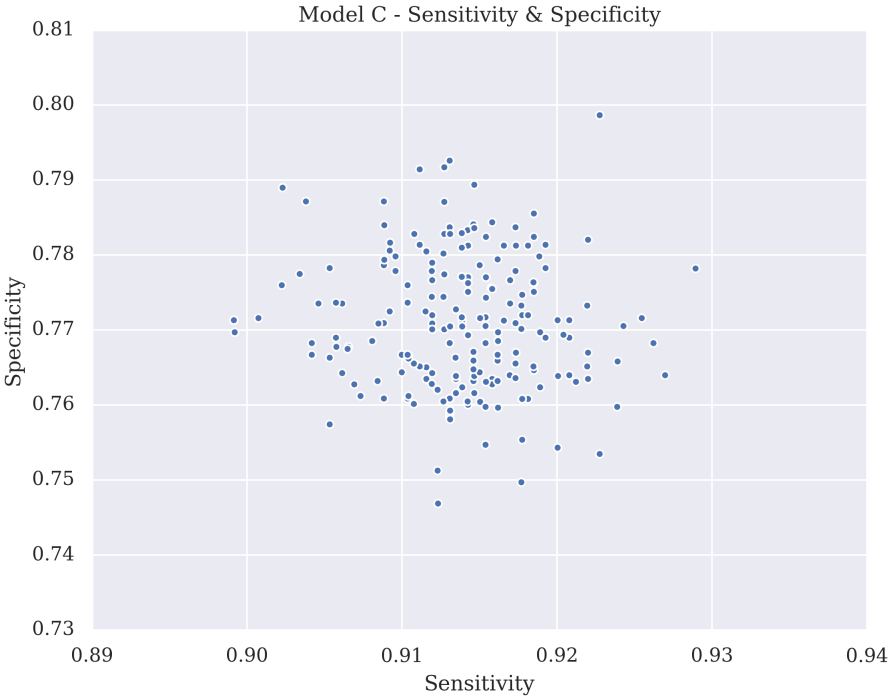
Obr. D.3: Výsledky křížové validace pro Model C



Obr. D.4: Rozložení metrik **Sensitivity** a **Specificity** pro Model A



Obr. D.5: Rozložení metrik **Sensitivity** a **Specificity** pro Model B

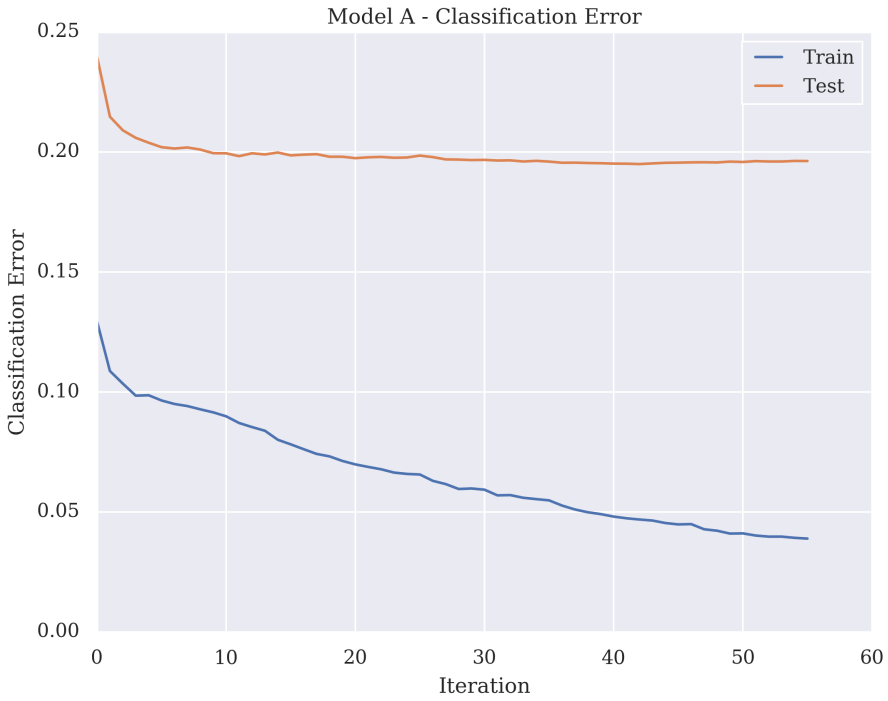


Obr. D.6: Rozložení metrik **Sensitivity** a **Specificity** pro Model C

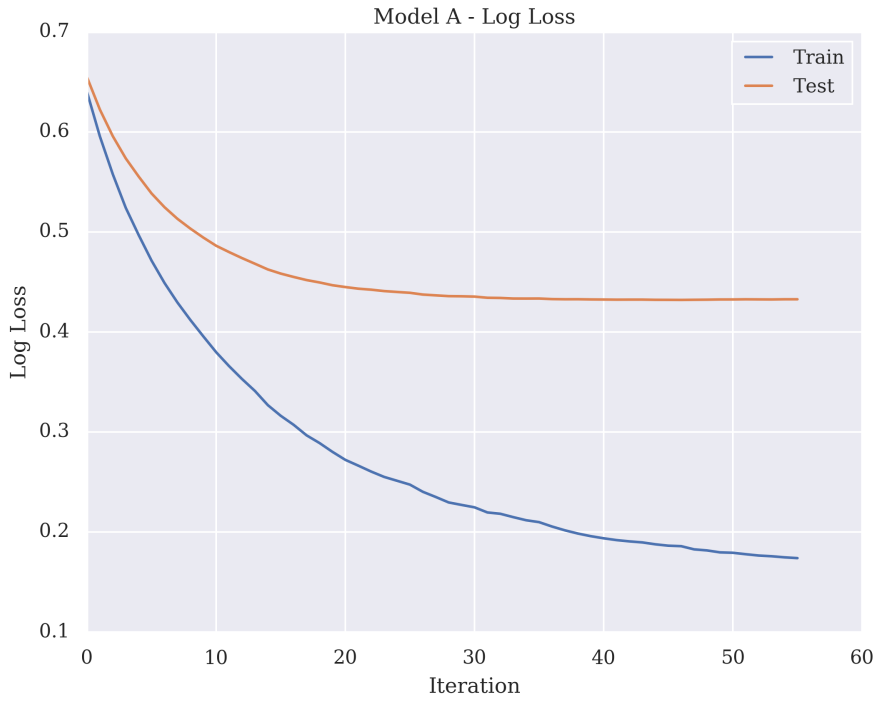
E Výsledky trénování modelu

Na obrázcích jsou zobrazeny jednotlivé evaluační metriky a jejich vývoj v rámci procesu trénování. Nejprve jsou zobrazeny metriky pro model A, poté pro model B a nakonec pro model C. Metriky začínají metrikou `error`, poté pokračují metrikou `logloss` a na závěr je uvedena metrika `auc`. Metrika `auc` byla zvolena jako metrika k zastavení procesu učení. Na grafech si lze povšimnout, že v případě modelu A je process ukončen po přibližně 60. iteraci, kdy již metrika `auc` dosahuje svého maxima a dále neroste. U Modelu B je využito plných 100 iterací, u Modelu C pouhých 20 iterací. Příslušným chováním by mělo být zabráněno procesu přeučení modelu na trénovacích datech.

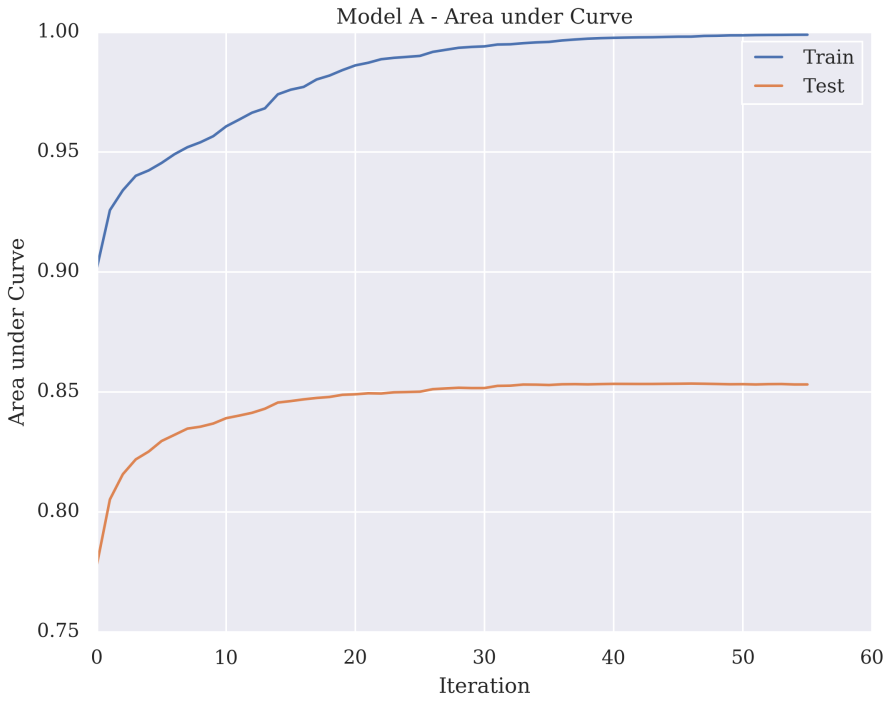
Dále pak následují grafy 10 nejdůležitějších parametrů pro každý z modelů. Z grafů lze vyčíst jak je který parametr pro daný model důležitý, a nebo například jak moc se model rozhoduje pouze na základě jednoho či více parametrů.



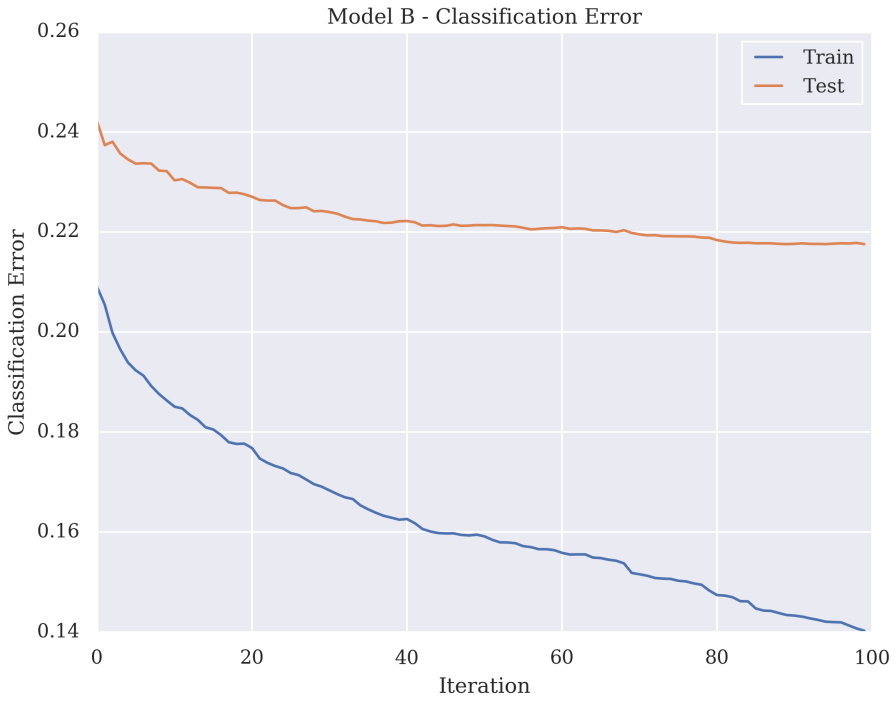
Obr. E.1: Metrika `error` v procesu trénování Modelu A



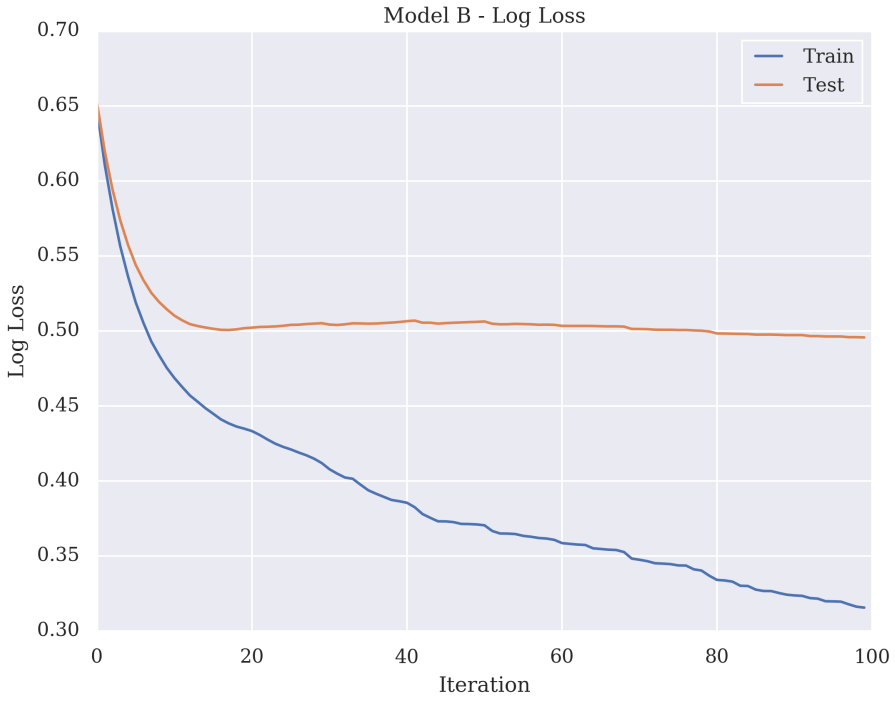
Obr. E.2: Metrika `logloss` v procesu trénování Modelu A



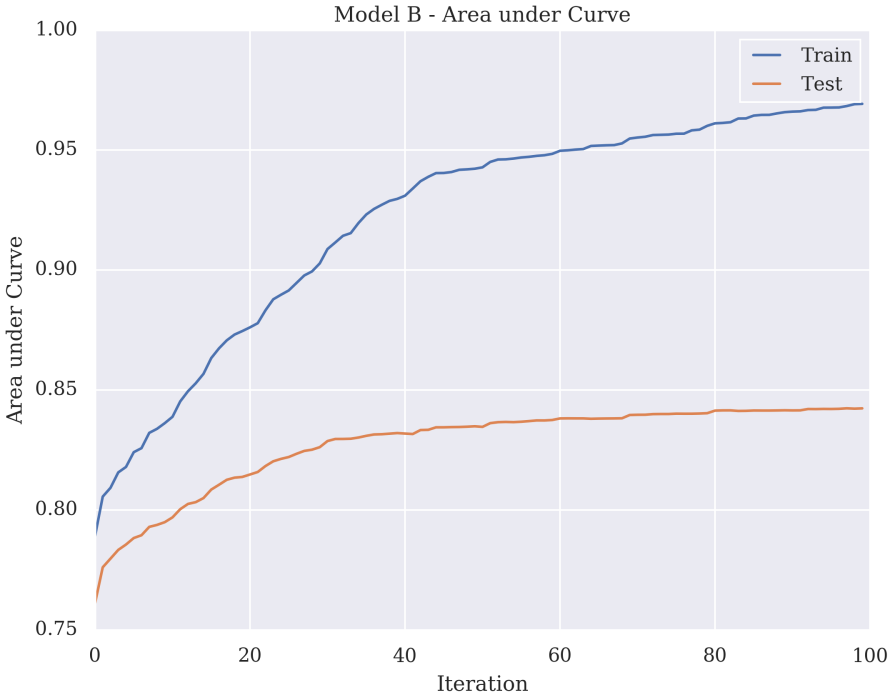
Obr. E.3: Metrika `auc` v procesu trénování Modelu A



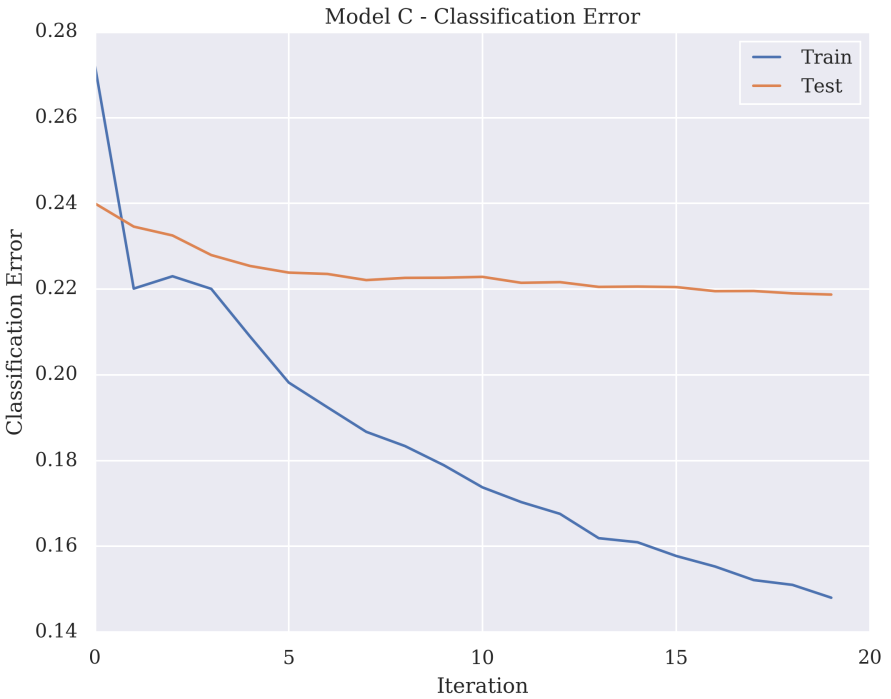
Obr. E.4: Metrika `error` v procesu trénování Modelu B



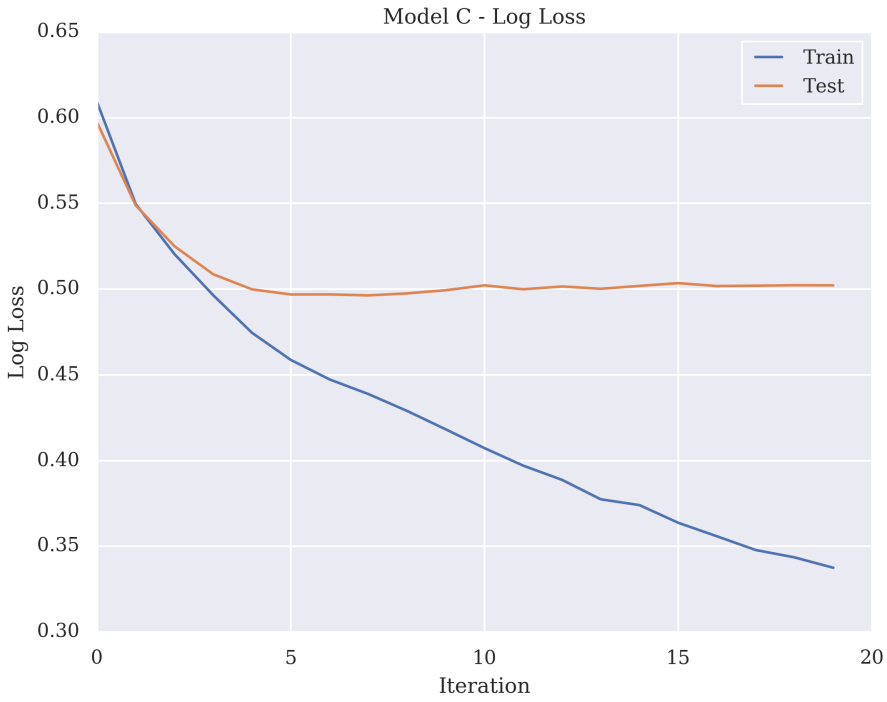
Obr. E.5: Metrika `logloss` v procesu trénování Modelu B



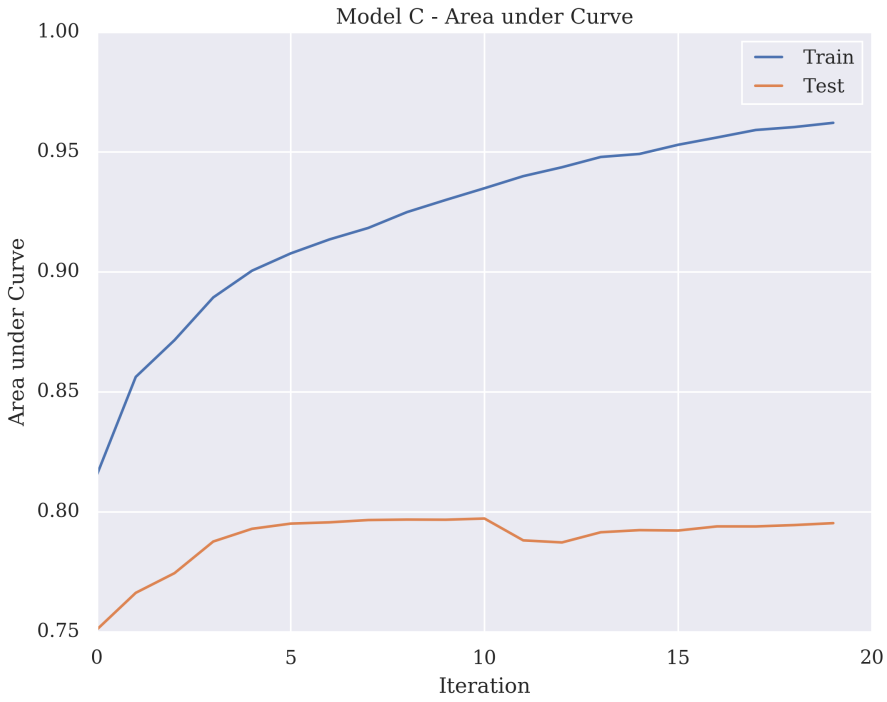
Obr. E.6: Metrika `auc` v procesu trénování Modelu B



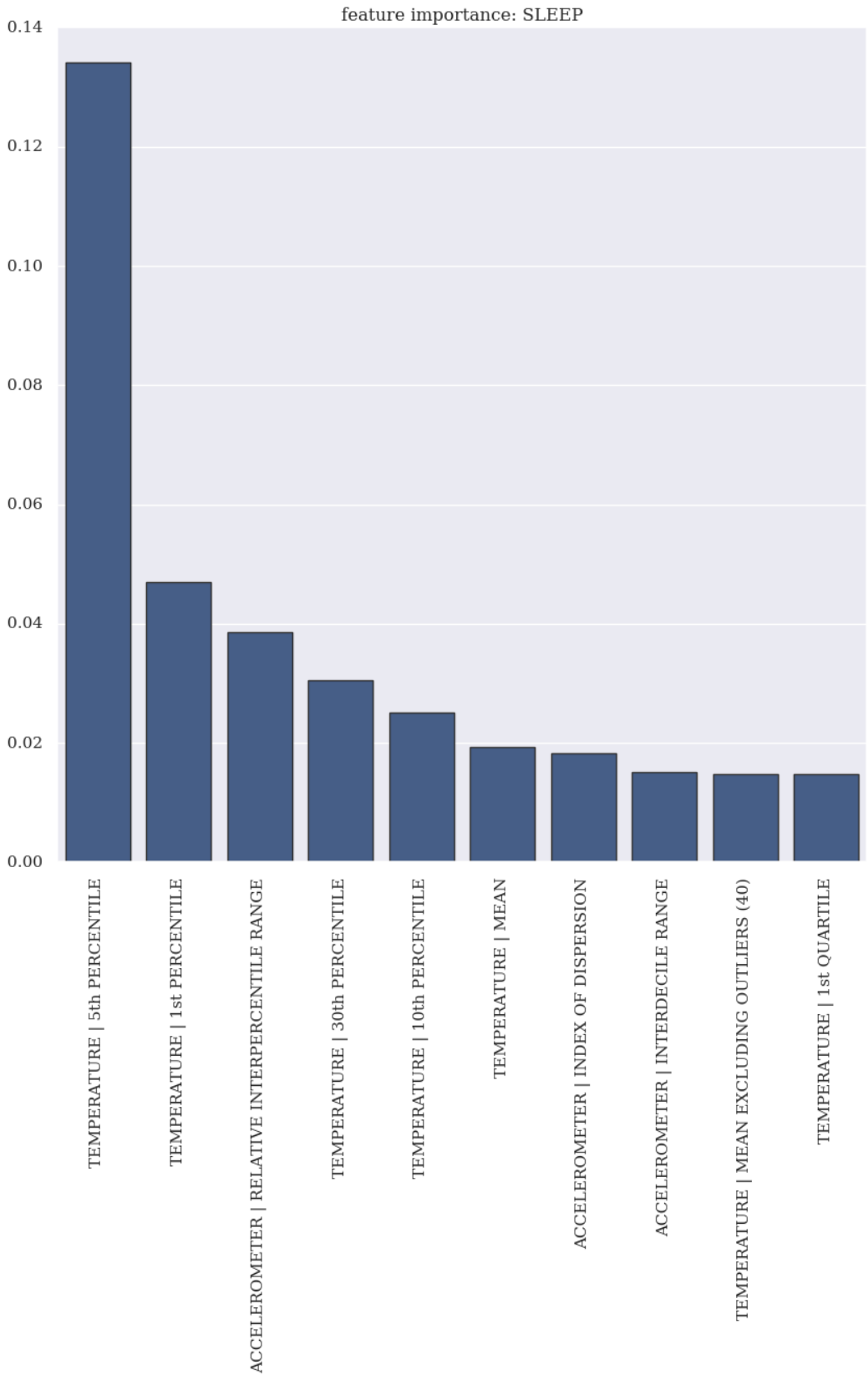
Obr. E.7: Metrika `error` v procesu trénování Modelu C



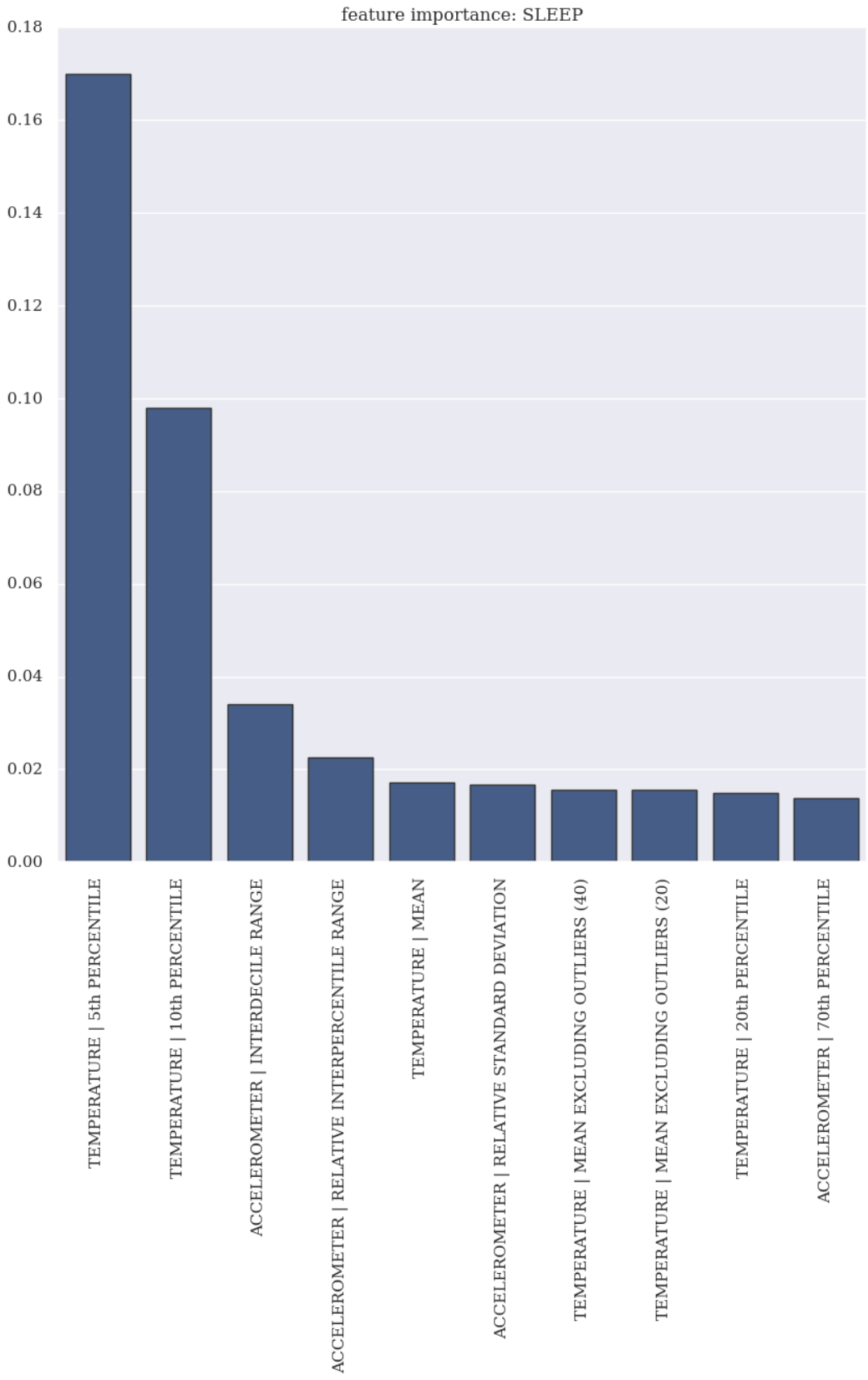
Obr. E.8: Metrika `logloss` v procesu trénování Modelu C



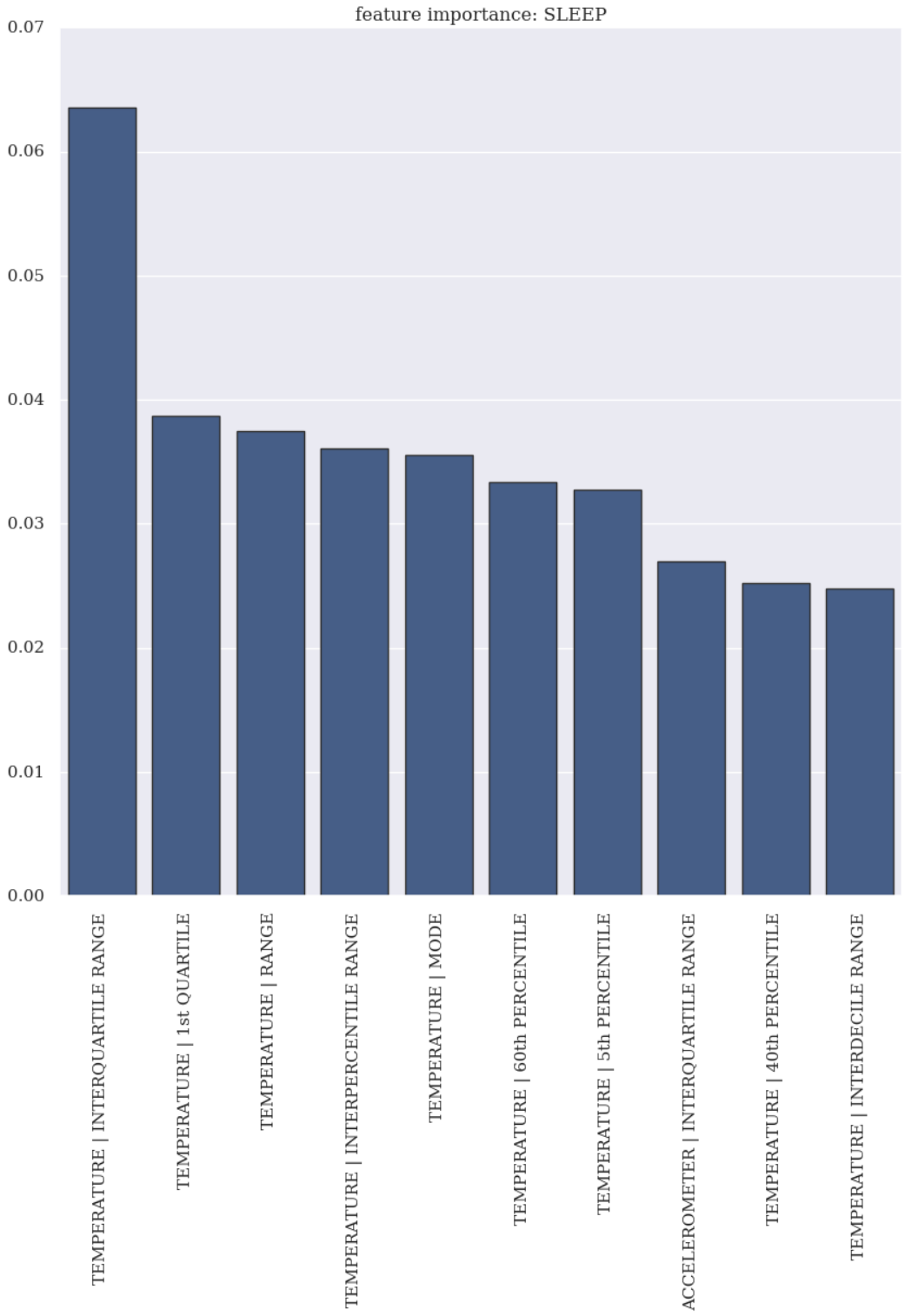
Obr. E.9: Metrika `auc` v procesu trénování Modelu B



Obr. E.10: Váha deseti nejdůležitějších parametrů pro Model A



Obr. E.11: Váha deseti nejdůležitějších parametrů pro Model B



Obr. E.12: Váha deseti nejdůležitějších parametrů pro Model C

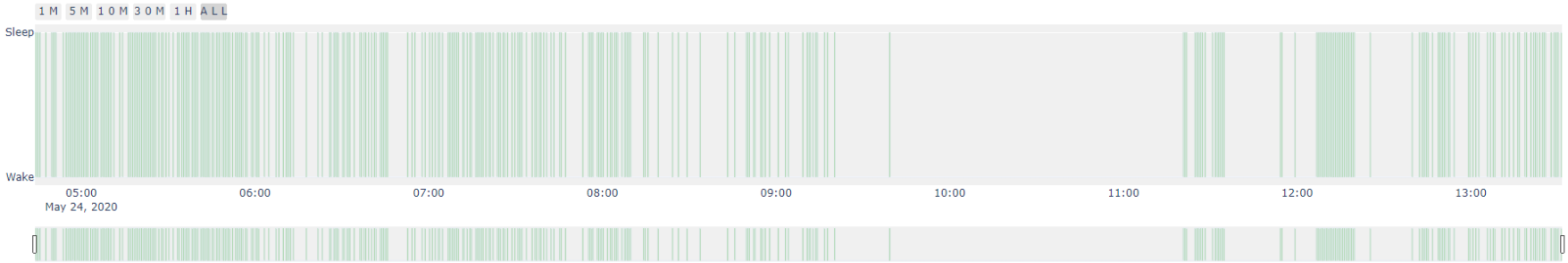
F Spánkový deník

V rámci testování systému byla provedena simulace testovací studie. Spánkový deník, který byl vytvořen vedoucím práce Ing. Jiřím Mekyskou, Ph.D. v rámci projektu NU20-04-00294, by byl využit při zamýšlené pilotní studii. Deník, společně se systémem, byl otestován přímo mnou, samotná pilotní studie proběhne v rámci projektu NU20-04-00294 po odevzdání diplomové práce.

Spánkový deník a výsledky systému oproti spánkovému deníku jsou obsaženy na následujících stránkách.

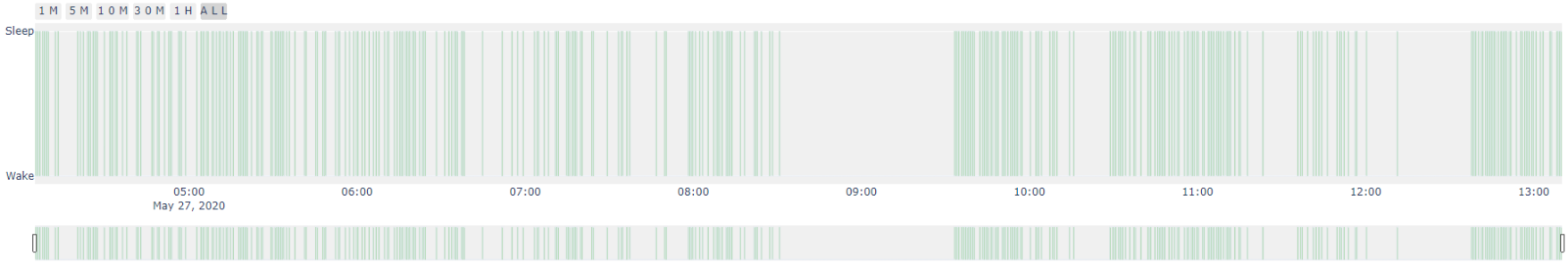
Body location: Left wrist | Creation date: 2020-05-23 | Description: Day 1

Body location: Left wrist | Creation date: 2020-05-24 | Description: Day 2

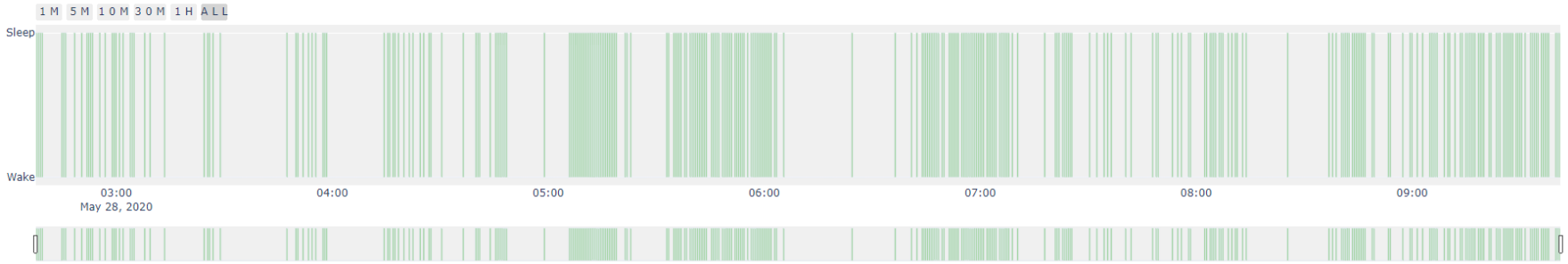


Body location: Left wrist | Creation date: 2020-05-25 | Description: Day 3

Body location: Left wrist | Creation date: 2020-05-27 | Description: Day 5



Body location: Left wrist | Creation date: 2020-05-31 | Description: Day 6



Body location: Left wrist | Creation date: 2020-05-31 | Description: Day 7

Body location: Left wrist | Creation date: 2020-05-31 | Description: Day 8

Body location: Left wrist | Creation date: 2020-05-31 | Description: Day 9

Spánkový deník (ver. 2020/05/06)

Jméno a příjmení: Marek Mikulec
Datum narození: 21. 11. 1995
ID: ME

vyplnit před spaním

vyplnit ráno (do 1 hod. po opuštění postele)

Datum	Příklad 2.5.2020	23.5.2020	24.5.2020	25.5.2020	27.5.2020	28.5.2020	29.5.2020	30.5.2020
1a. Kolikrát jste si šel/šla během dne lehnout? 1b. Kolik minut jste během dne prospal/a?	2x 40 min.	0x 0 min.	0x 0 min.	0x 0 min.	0x 0 min.	0x 0 min.	0x 0 min.	0x 0 min.
2a. Kolik alkoholických nápojů jste přes den vypil/a? 2b. V kolik hodin jste vypil/a poslední?	1 20:15	0	0	0	0	0	1 19:00	1 17:10
3a. Kolik nápojů s kofeinem (kafe, čaj, energy drinků) jste přes den vypil/a? 3b. V kolik hodin jste vypil/a poslední?	3 17:32	2 19:00	1 19:00	1 19:20	0	0	1 21:00	1 23:00
4. Užil/a jste lék na spaní? (pokud ano jaký, kolik a v jaký čas)		Ne	Ne	Ne	Ne	Ne	Ne	Ne
5. V kolik hodin jste ulehl/a do postele?	22:35	1:46	4:47	0:35	4:05	2:41	6:23	3:25
6. V kolik hodin jste se snažil/a usnout?	22:50	2:20	5:34	0:43	6:24	3:10	6:34	3:40
7. Jak dlouho vám přibližně trvalo usnout? (nesledujte při usínání hodiny)	10 min.	2 min.	10 min.	10 min.	10 min.	5 min.	2 min.	2 min.
8. Probuzení během noci (pro každé probuzení: přibližně od kdy – do kdy)	1:12 – 1:17 3:45 – 3:58	9:00-9:05 9:10-9:15 9:20-9:25	6:06-8:10	2:30-2:50 3:10-3:20 4:00-4:20 6:00-6:10 7:00-7:30	7:30-8:40	9:00-9:10	9:00-9:02 9:15-9:20	5:00-5:02 8:00-8:05
9. V kolik hodin jste se probudil/a?	7:12	9:30	12:55	9:20	12:30	9:40	10:00	10:44
10. V kolik hodin jste opustil/a postel?	7:20	10:00	13:10	9:25	12:34	9:43	10:22	10:45
11. Jak dobře jste se vyspal/a? (1–5, viz legenda pod tabulkou)	2	3	5	2	4	3	4	2
12. Jak svěže jste se po probuzení cítil/a? (1–5, viz legenda pod tabulkou)	3	3	5	2	3	3	4	3
13. Komentář (nepovinné)	nachlazení			Cílené probuzení	26.5. beze spánku			

Hodnocení: 1 = velmi dobře, 2 = dobře, 3 = průměrně, 4 = špatně, 5 = velmi špatně

Instrukce k deníku a jeho položkám

Jak často deník vyplňovat? Prosím vyplňujte denně před i po spánku dle vyznačených polí v deníku.

Co dělat, když vynechám den? Snažte se vyplňovat 7 dní v kuse. Pokud vynecháte, nic nevyplňujte a pokračujte dalším dnem. Musí být 7 záznamů.

Co dělat, když něco nestandardního (třeba nachlazení/chřipka) ovlivní můj spánek? Informujte o tomto v poli č. 13.

Jak přesné musí být uvedené časy? Uvádějte přibližné odhady, nesledujte zbytečně hodiny (např. při usínání).

2a. Kolik alkoholických nápojů jste přes den vypil/a? Za 1 alkoholický nápoj se považuje 0,5l piva, 2dl vína, nebo 1 panák tvrdého alkoholu.

5. V kolik hodin jste ulehl/a do postele? Toto nemusí být čas, kdy se začnete snažit spát. Např. si budete ještě číst.

6. V kolik hodin jste se snažil/a usnout? Toto je opravdu čas, kdy jste se odebrali ke spánku.

9. V kolik hodin jste se probudil/a? Uveďte čas posledního probuzení, po kterém jste ráno již neusnul/a.

10. V kolik hodin jste opustil/a postel? Mezi časem probuzení a tímto časem jste mohli např. jen polehávat, brouzdat na telefonu atd.

11. Jak dobře jste se vyspal/a? Zhodnoťte, jestli byl váš spánek dobrý, špatný…

12. Jak svěže jste se po probuzení cítil/a? Jak jste se cítil/a během pár minut po probuzení?

Instrukce k náramku (aktigrafu)

Náramek bude zaznamenávat Váš pohyb nedominantní ruky, teplotu a míru okolního osvětlení. Žádné další informace nejsou zaznamenávány.

Náramek se nenabíjí. Než Vám byl předán, byl dostatečně nabit.

Nenamáčejte náramek do vody.

Náramek se nasazuje na nedominantní ruku. Pokud jste pravák/pravačka, dejte si náramek na ruku levou. Pokud jste levák/levačka, dejte si náramek na ruku pravou.

Náramek si na nedominantní ruku nasaďte vždy před tím, něž ulehnete do postele. Sundejte ho hned po tom, co ráno postel opustíte.

Náramek si upevněte tak, aby vás příliš nestahoval, ale **aby také nebyl na ruce příliš volný**.

Kontakt pro dotazy a řešení technických potíží

MUDr. Ivona Morávková, +420 XXX XXX XXX

Děkujeme Vám za spolupráci. Vaše trpělivost při vyplňování deníku a nošení náramku pomůže při výzkumu a vývoji nové metody diagnózy onemocnění s Lewyho tělísky v raném stádiu. Ta mj. umožní zvýšit kvalitu života pacientů s těmito onemocněními.